THE STOCHASTIC HARMONIC AUTOREGRESSIVE PARAMETRIC (SHARP) WEATHER GENERATOR

by

Kimberly L. Smith

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Atmospheric Sciences

The University of Utah

August 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Kimberly L. Smith**
has been approved by the following supervisory committee members:

Courtenay Strong ,	Chair(s)	**24 May 2017** Date Approved
Firas Rassoul-Agha ,	Member	**24 May 2017** Date Approved
John Horel ,	Member	**24 May 2017** Date Approved
Thomas Reichler ,	Member	**24 May 2017** Date Approved
Andrew Wood ,	Member	**24 May 2017** Date Approved

by **Kevin Perry** , Chair/Dean of
the Department/College/School of **Atmospheric Sciences**
and by **David B. Kieda** , Dean of The Graduate School.

ABSTRACT

Stochastic weather generators (SWGs) are statistically-based point-scale models of meteorological data that are driven by random number generators. Commonly taking observational data or low-resolution global climate model data as input, they are useful tools for generating many realizations of possible climate scenarios for use in impacts studies. This dissertation presents the stochastic harmonic autoregressive parametric (SHArP) weather generator. SHArP is based on previous SWGs but it generates air temperature values directly instead of prescribing and removing the mean and standard deviations in advance and generating temperature residuals. In addition, in both the precipitation process and the temperature process, SHArP includes nonstationarity due to oceanic modes of variability. During frontal passage, the precipitation-responsive autocorrelated transitions result in more realistic temperatures. The multisite generalization of SHArP presents a challenge due to an exponential increase in the number of noise coefficient matrices as the number of sites increases, but empirical orthogonal function analysis is applied to the precipitation patterns over the domain in order to reduce the number of noise coefficient matrices to a reasonable number. For multisite precipitation simulation, a trend due to climate change is added. Even though they are statistically-based, SWGs are limited in their ability to capture meteorological extremes, including dry and wet spells. The second-order Markovian probabilities of precipitation at a single site are modified using the method of large deviations. This mathematically-based method is shown to accurately modify the probabilities of precipitation to produce binary precipitation occurrence time series that are extreme yet statistically consistent with the input data without needing to "wait to get lucky" for those extreme events to occur in very long simulations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grants EPS-1135482, EPS-1135483, EPS-1208732, and DMS-1407574. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Provision of computer infrastructure by the Center for High Performance Computing at the University of Utah is gratefully acknowledged. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Travel support from the Global Change and Sustainability Center (GCSC) at the University of Utah is also gratefully acknowledged.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The Great Basin, which is made up of many smaller basins including the Great Salt Lake Basin, is facing a future with highly variable water availability due to the changing climate. A rise in air temperatures will lead to less precipitation falling as snow in the winter months as well as earlier runoff timing, which is problematic for stakeholders who depend on wintertime precipitation to use throughout the dry summer months. Global climate models (GCMs) are useful tools for studying how the climate is changing on continental scales; however, their resolution is too coarse to resolve impacts on regional scales, especially in areas of complex terrain like the Great Basin. GCMs also have a difficult time capturing the low-frequency connections between the Pacific Ocean and the Great Basin (Smith et al., 2015). The performance of state-of-the-art GCMs has been evaluated in terms of ability to capture the "extremes" in precipitation and temperature, and it has been found that GCMs do not have the skill to capture the extremes, though they perform better at temperature extremes than precipitation (Kiktev et al., 2007). In addition, even when the GCMs are able to capture extremes moderately well, their computational expense allows only a small number of model runs. To circumvent these limitations, a nonstationary, daily stochastic weather generator is introduced to realistically capture the precipitation and temperature trends within the Great Salt Lake Basin as well as the meteorological extremes that play an important role in how climate change will impact the region.

Statistically-based weather models referred to as "stochastic weather generators" (SWGs) use mathematical formulations driven by random number generators to produce precipitation and other nonprecipitation variables that match the statistical properties of the training data at a given location (Wilks and Wilby,

1999). SWGs commonly take either observational meteorological data or output from GCMs as input. Because SWGs work on a point-scale, they can additionally provide a downscaling of the low-resolution GCM output. The earliest known stochastic weather generators (SWGs) were essentially precipitation simulators only (e.g., Gabriel and Neumann, 1962); generating precipitation is a natural first step since the presence (or lack) of precipitation often affects nonprecipitation variables such as air temperature. They have since become more elaborate and can now generate variables including air temperature and solar radiation in addition to precipitation occurrence and amount (e.g., Matalas, 1967; Richardson, 1981).

The first SWG-related studies generated precipitation occurrence using a two-state, first-order Markov chain framework (Bailey, 1964; Richardson, 1981; Roldàn and Woolhiser, 1982), meaning that the probability of precipitation occurrence on a given day is only dependent on whether precipitation occurred on the previous day. Precipitation amount was modeled separately, and later, the binary precipitation states were used to determine models for generating air temperature and solar radiation. Other SWG studies considered instead a two-state, second-order Markov chain framework (e.g., Stern and Coe, 1984; Wilks, 1999a). Markov chains of higher order are better able to capture dry spells than first-order Markov chains, and are thus more useful in studies that involve the areas of the western U.S. where dry spells are common, such as the semi-arid Great Basin.

Wilks (1998) introduced the widely known multisite generalization model of precipitation occurrence and amount based on chain-dependent processes: a two-state, second-order Markov chain for occurrence and a mixed exponential distribution for amount. These methods were first described in Matalas (1967) and Todorovic and Woolhiser (1975) and later applied in Richardson (1981). This is done by applying spatially correlated yet time-independent random vectors on the models of each individual site within the domain (Wilks, 1998). With this method, each site retains its own statistical properties while maintaining realistic spatial correlations between sites. Wilks (1999b) proposed a treatment of spatial correlation for complex terrain, incorporating solar radiation, precipitation, and temperature. The spatial correlations used for the multisite generalization are

functions of both horizontal and vertical distance between sites so the relationships between mountain locations and valley locations remained realistic.

In addition to the well-known and widely-used parametric weather generators, a multitude of nonparametric SWGs and generalized linear models (GLMs) have been developed. These data-driven SWGs involve either kernel density estimation (e.g., Rajagopalan et al., 1997; Harrold et al., 2003) or resampling via *k*-nearest neighbor (k-NN) bootstrapping (e.g., Rajagopalan and Lall, 1999; Caraway et al., 2014). These models are an alternative to the linear models presented in the parametric SWGs, which are unable to capture the nonlinear relationships between meteorological variables and rely on statistical relationships that GLMs do not. Stern and Coe (1984) first introduced using GLMs in SWGs, which has also been increasing in popularity because they can easily model discrete variables and variables with non-normal distributions (Furrer and Katz, 2007). In addition, GLMs have the ability to treat ENSO and other major oceanic modes of variability as continuous variables (e.g., Chandler, 2005). McCullagh and Nelder (1989) offer more comprehensive details on GLMs.

Most studies involving SWGs largely follow the precipitation methods introduced in Matalas (1967) and air temperature methods introduced in Richardson (1981). The Richardson method of generating stochastic temperature involves prescribing and then removing the mean and standard deviations in advance, and then simulating only the temperature residuals. A limitation of the widely-used Richardson model is that its mean and standard deviation switch abruptly between wet- and dry-state values prescribed in advance of the simulation, and temperature is not simulated directly but rather through its residuals. Instead, in reality, there are smooth, autocorrelated transitions between wet- and dry-state values. The goal of this research was to introduce a linear model that simulates temperature values directly.

This dissertation introduces the stochastic harmonic autoregressive parametric (SHArP) weather generator, which uses a linear model for generating stochastic temperature values directly rather than prescribing the mean and standard deviation in advance and generating temperature residuals as is done in the Richardson

method. The fidelity of the model is established using only maximum temperature at a single site, and then the model is generalized to maximum and minimum temperature at multiple sites across a region of complex terrain. The method of large deviations is shown to capture extremes in the two-stae Markov chain used in SHArP. In addition, an existing precipitation model compatible with SHArP is detailed and extended to include trends associated with climate change.

This dissertation consists of one published manuscript (Chapter 2), one submitted manuscript (Chapter 3), and one manuscript in preparation for publication (Chapter 4). The rest of this chapter includes two sections on the existing stochastic precipitation and air temperature frameworks, which provided starting points for the research presented here. Chapter 2 (Smith et al., 2017) introduces the mathematics behind the method for generating stochastic simulations of air temperature directly with the SHArP framework and illustrates how this new approach produces more realistic temporal evolution of temperature than the existing Richardson method. Chapter 3 describes the process for generalizing SHArP to multiple sites via empirical orthogonal function (EOF) analysis of the precipitation patterns over the simulation domain, and has been submitted to a journal for peer review. Chapter 4 shows how the mathematical method of large deviations can be used to efficiently generate extreme realizations of precipitation (e.g., prolonged drought) directly rather than waiting for such events to arise randomly in very long simulations. This chapter is in preparation for submission to a peer reviewed journal.

1.2 Existing Precipitation Occurrence Framework

To define whether precipitation occurred, we consider the precipitation amount y_t at each location, where the subscript t counts days $t = 1, 2, \ldots, 365Y$ for a record of Y years. The indicator variable

$$X_t = \begin{cases} 1, & \text{if } y_t \geq h; \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

takes the value 1 to indicate a "wet" day (precipitation of at least $h = 0.25$ mm) and takes the value 0 to indicate a dry day. In the second-order Markov chain

framework (e.g., Stern and Coe, 1984; Wilks, 1999a), the probability of observing a dry day depends on the sequence of wet or dry conditions occurring during the preceding two days

$$p_{ij0}(t) = P\left\{X_t = 0 | X_{t-1} = j,\ X_{t-2} = i\right\};\ \ t = 1, 2, \ldots, 365Y. \tag{1.2}$$

Equation (1.2) represents four probabilities (one for each of the possible sequences of wet or dry on days $t-1$ and $t-2$). For example, $p_{010}(t)$ denotes the probability of observing the sequence dry-wet-dry completing on day t. The four remaining p terms represent the probability of observing a wet day following a given sequence of wet or dry days given by $p_{ij1}(t) = 1 - p_{ij0}(t)$.

The likelihood of observing a particular sequence of X_t is given by the Bernoulli measure (Klenke, 2013)

$$L = \prod_{i,j=0}^{1} \prod_{t=1}^{365Y} p_{ij0}(t)^{b_{ij0}(t)} \left(1 - p_{ij0}(t)\right)^{b_{ij1}(t)}, \tag{1.3}$$

where

$$b_{ijk}(t) = \begin{cases} 1, & \text{if } X_{t-2} = i,\ X_{t-1} = j,\ X_t = k; \\ 0, & \text{otherwise.} \end{cases} \tag{1.4}$$

Usually, the Markov chain process is applied assuming stationarity. Cyclostationarity indicates that the p_{ij0} terms are periodic, meaning $p_{ij0}(t + K365) = p_{ij0}(t)$ for any integer K. We can then rewrite the product over t in (1.3) as a product over day of year n

$$L = \prod_{i,j=0}^{1} \prod_{n=1}^{365} p_{ij0}(n)^{N_{ij0}(n)} \left(1 - p_{ij0}(n)\right)^{N_{ij1}(n)}, \tag{1.5}$$

where N_{ijk} is the number of times that the sequence $\{X_{t-2} = i, X_{t-1} = j, X_t = k\}$ occurred on day of year n.

We seek to maximize L, or equivalently, to maximize its natural log

$$\ln L = \sum_{i,j=0}^{1} \sum_{n-1}^{365} \left[N_{ij0}(n) \ln p_{ij0}(n) + N_{ij1}(n) \ln(1 - p_{ij0}(n))\right]. \tag{1.6}$$

It is convenient to write $p_{ij0}(n)$ as an inverse logit

$$p_{ij0}(n) = \text{logit}^{-1}\left[G_{ij0}(n)\right] = \frac{\exp(G_{ij0}(n))}{1 + \exp(G_{ij0}(n))}, \tag{1.7}$$

allowing an unconstrained maximization of $\ln L$ with respect to $-\infty < G_{ij0}(n) < \infty$, instead of a constrained maximization with respect to $0 \leq p_{ij0}(n) \leq 1$.

For simplicity, we represent seasonality via a harmonic formulation for $G_{ij0}(n)$

$$G_{ij0}(n) = \sum_{k=1}^{m_{ij0}} a_{ij0}^{(k)} \phi^{(k)} \tag{1.8}$$

$$\phi^{(k)} = \begin{cases} \cos\left[(k-1)\pi n/365\right], & k = 1,3,5,\ldots \\ \sin(k\pi n/365), & k = 2,4,6,\ldots \end{cases} \tag{1.9}$$

We use the first and second derivatives of $\ln L$

$$\frac{\partial \ln L}{\partial a_{ij0}^{(k)}} = \sum_{i,j=0}^{1} \sum_{n=1}^{365} \left\{ \left[N_{ij0}(n) - (N_{ij0}(n) + N_{ij1}(n)) \frac{\exp(G_{ij0}(n))}{1+\exp(G_{ij0}(n)} \right] \phi_{ij0}^{(k)} \right\}, \tag{1.10}$$

$$\frac{\partial^2 \ln L}{\partial a_{ij0}^{(k)} \partial a_{ij0}^{(l)}} = \sum_{i,j=0}^{1} \sum_{n=1}^{365} \left[-(N_{ij0}(n) + N_{ij1}(n)) \frac{\exp(G_{ij0}(n)}{\left[1+\exp(G_{ij0}(n))\right]^2} \phi_{ij0}^{(k)} \phi_{ij0}^{(l)} \right] \tag{1.11}$$

with a Newton-Raphson iterative procedure to maximize $\ln L$ with respect to the a_{ij0} parameters.

A description of the existing precipitation occurrence and amount frameworks for multiple sites is presented as part of Chapter 3.

1.3 Existing Air Temperature Framework

The Richardson stochastic temperature method (Richardson, 1981) begins by computing the residuals of the variables of interest (e.g., maximum air temperature). The standardized residuals are computed by subtracting the mean and dividing by the standard deviation

$$\chi_{p,i}(j) = \begin{cases} \frac{X_{p,i}(j) - \bar{X}_i^0(j)}{\sigma_i^0(j)}, & \text{if } Y_{p,i} = 0; \\ \frac{X_{p,i}(j) - \bar{X}_i^1(j)}{\sigma_i^1(j)}, & \text{if } Y_{p,i} > 0, \end{cases} \tag{1.12}$$

where $\bar{X}_i^0(j)$ and $\sigma_i^0(j)$ are the mean and standard deviation for a dry day ($Y_{p,i} = 0$), $\bar{X}_i^1(j)$ and $\sigma_i^1(j)$ are the mean and standard deviation for a wet day ($Y_{p,i} > 0$), and $\chi_{p,i}(j)$ is the residual component for variable j.

Following Matalas (1967), the multivariate generation model for generating residual series of temperature and solar radiation is

$$\chi_{p,i}(j) = A\chi_{p,i-1}(j) + B\epsilon_{p,i}(j), \tag{1.13}$$

where $\chi_{p,i}(j)$ and $\chi_{p,i-1}(j)$ are 3×1 matrices for days i and $i-1$ of year p whose elements are residuals of maximum temperature ($j = 1$), minimum temperature ($j = 2$), and solar radiation ($j = 3$). $\epsilon_{p,i}(j)$ is a 3×1 matrix of independent random components that are normally distributed.

The A and B matrices are determined from

$$A = M_1 M_0^{-1} \quad \text{and} \tag{1.14}$$

$$BB^T = M_0 - M_1 M_0^{-1} M_1^T, \tag{1.15}$$

where the superscripts -1 and T denote the inverse and transpose of the matrix, respectively, and M_0 and M_1 are the lag 0 and lag 1 covariance matrices. Principal component analysis can be used to solve for B in (1.15):

$$B^T B = \lambda, \tag{1.16}$$

where λ is an 3×3 diagonal matrix whose elements are the eigenvalues of $M_0 - M_1 M_0^{-1} M_1^T$ (Matalas, 1967). Because the $\chi_{p,i}(j)$ series have unity variances, the M_0 and M_1 matrices contain the lag 0 and lag 1 cross-correlation coefficients, which may be written as

$$M_0 = \begin{bmatrix} 1 & \rho_0(1,2) & \rho_0(1,3) \\ \rho_0(2,1) & 1 & \rho_0(2,3) \\ \rho_0(3,1) & \rho_0(3,2) & 1 \end{bmatrix},$$

$$M_1 = \begin{bmatrix} \rho_1(1,1) & \rho_1(1,2) & \rho_1(1,3) \\ \rho_1(2,1) & \rho_1(2,2) & \rho_1(2,3) \\ \rho_1(3,1) & \rho_1(3,2) & \rho_1(3,3) \end{bmatrix}$$

where $\rho_0(j,k)$ is the lag 0 cross-correlation coefficient between variables j and k, $\rho_1(j,k)$ is the cross-correlation coefficient between variables j and k with variable k lagged 1 day in relation to variable j, and $\rho_1(j)$ is the lag 1 serial correlation for variable j. Since $\rho_0(j,k) = \rho_0(k,j)$, M_0 is a symmetric matrix.

To compute the complete synthetic series of maximum temperature, minimum temperature, and solar radiation, the means and standard deviations that were removed to create the original residual series are then added or multiplied back to the synthetic residuals.

1.4 References

Bailey, N. T. J., 1964: *The Elements of Stochastic Processes*. John Wiley, New York, 39 pp.

Caraway, N. M., J. L. McCreight, and B. Rajagopalan, 2014: Multisite stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *Journal of Hydrology*, **508**, 197 – 213, doi:http://dx.doi.org/10.1016/j.jhydrol.2013.10.054, URL http://www.sciencedirect.com/science/article/pii/S0022169413007981.

Chandler, R. E., 2005: On the use of generalized linear models for interpreting climate variability. *Environmetrics*, **16 (7)**, 699–715, doi:10.1002/env.731, URL http://dx.doi.org/10.1002/env.731.

Furrer, E. M. and R. W. Katz, 2007: Generalized linear modeling approach to stochastic weather generators. *Climate Research*, **34 (2)**, 129–144, URL http://www.int-res.com/abstracts/cr/v34/n2/p129-144/.

Gabriel, K. R. and J. Neumann, 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, **88 (375)**, 90–95, doi:10.1002/qj.49708837511, URL http://dx.doi.org/10.1002/qj.49708837511.

Harrold, T. I., A. Sharma, and S. J. Sheather, 2003: A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, **39 (12)**, n/a–n/a, doi:10.1029/2003WR002570, URL http://dx.doi.org/10.1029/2003WR002570, 1343.

Kiktev, D., J. Caesar, L. V. Alexander, H. Shiogama, and M. Collier, 2007: Comparison of observed and multimodeled trends in annual extremes of temperature and precipitation. *Geophysical Research Letters*, **34 (10)**, n/a–n/a, doi:10.1029/2007GL029539, URL http://dx.doi.org/10.1029/2007GL029539.

Klenke, A., 2013: *Probability Theory: A Comprehensive Course*. 2nd ed., Springer, Heidelberg, Germany.

Matalas, N. C., 1967: Mathematical assessment of synthetic hydrology. *Water Resources Research*, **3 (4)**, 937–945, doi:10.1029/WR003i004p00937, URL http://dx.doi.org/10.1029/WR003i004p00937.

McCullagh, P. and J. Nelder, 1989: *Generalized Linear Models*. 2nd ed., Chapman & Hall, London.

Rajagopalan, B. and U. Lall, 1999: A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, **35 (10)**, 3089–3101, doi:10.1029/1999WR900028, URL http://dx.doi.org/10.1029/1999WR900028.

Rajagopalan, B., U. Lall, and D. G. Tarboton, 1997: Evaluation of kernel density estimation methods for daily precipitation resampling. *Stochastic Hydrology and Hydraulics*, **11 (6)**, 523–547, doi:10.1007/BF02428432, URL http://dx.doi.org/10.1007/BF02428432.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17 (1)**, 182–190, doi:10.1029/WR017i001p00182, URL http://dx.doi.org/10.1029/WR017i001p00182.

Roldàn, J. and D. A. Woolhiser, 1982: Stochastic daily precipitation models: 1. A comparison of occurrence processes. *Water Resources Research*, **18 (5)**, 1451–1459.

Smith, K., C. Strong, and F. Rassoul-Agha, 2017: A new method for generating stochastic simulations of daily air temperature for use in weather generators. *Journal of Applied Meteorology and Climatology*, **56 (4)**, 953–963, doi:10.1175/JAMC-D-16-0122.1, URL http://dx.doi.org/10.1175/JAMC-D-16-0122.1, http://dx.doi.org/10.1175/JAMC-D-16-0122.1.

Smith, K., C. Strong, and S.-Y. Wang, 2015: Connectivity between historical Great Basin precipitation and Pacific Ocean variability: A CMIP5 model evaluation. *Journal of Climate*, **28 (15)**, 6096–6112, doi:10.1175/JCLI-D-14-00488.1, URL http://dx.doi.org/10.1175/JCLI-D-14-00488.1.

Stern, R. D. and R. Coe, 1984: A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, **147 (1)**, 1–34, URL http://www.jstor.org/stable/2981736.

Todorovic, P. and D. A. Woolhiser, 1975: A stochastic model of ω-day precipitation. *Journal of Applied Meteorology*, **14**, 17–24, doi:10.1175/1520-0450(1975)014⟨0017:ASMODP⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0450(1975)014<0017:ASMODP>2.0.CO;2.

Wilks, D., 1998: Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, **210 (1–4)**, 178 – 191, doi:10.1016/S0022-1694(98)00186-3, URL http://www.sciencedirect.com/science/article/pii/S0022169498001863.

Wilks, D., 1999a: Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, **93 (3)**, 153–169, URL http://www.sciencedirect.com/science/article/pii/S0168192398001257.

Wilks, D., 1999b: Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, **96 (1–3)**, 85 – 101, doi:10.1016/S0168-1923(99)00037-4, URL http://www.sciencedirect.com/science/article/pii/S0168192399000374.

Wilks, D. S. and R. L. Wilby, 1999: The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, **23 (3)**, 329–357, doi:10.1177/030913339902300302, URL http://ppg.sagepub.com/content/23/3/329.abstract, http://ppg.sagepub.com/content/23/3/329.full.pdf+html.

CHAPTER 2

A NEW METHOD FOR GENERATING STOCHASTIC SIMULATIONS OF DAILY AIR TEMPERATURE FOR USE IN WEATHER GENERATORS[1]

2.1 Abstract

A stochastic harmonic autoregressive parametric (SHArP) weather generator is presented that simulates trended, nonstationary temperature values directly, circumventing the conventional approach of adding simulated standardized anomalies of temperature to a prescribed cyclostationary mean. The model mean makes autocorrelated transitions between wet- and dry-state values, and its parameters are determined by optimizing harmonic and trend terms. The precipitation-responsive autocorrelated transitions yield more realistic temperature behavior during frontal passage in comparison to prior models which switch abruptly between wet- and dry-state means. If the stochastic ("noise") term is assumed to have constant amplitude, analytical results are available via maximum likelihood estimation (MLE) and are equivalent to least squares estimation (LSE). Where observations motivate a seasonally-varying noise coefficient, MLE becomes nonlinear, and we formulate an analytical solution via LSE. For illustration, SHArP is shown to produce realistic representations of daily maximum air temperature at a single site, which for the study is the Salt Lake City International Airport (KSLC). SHArP reduces the temperature bias following frontal passages by over 2°C in three sea-

---

[1]Smith, Kimberly, Courtenay Strong, Firas Rassoul-Agha 2017: A new method for generating stochastic simulations of daily air temperature for use in weather generators, *J. Appl. Meteor. Climatol.*, **56 (4)**, 953–963, 10.1175/JAMC-D-16-0122.1.

sons. A method for generalizing the model to multiple variables at multiple sites is discussed.

2.2 Introduction

The drought-stricken western U.S., including the Great Basin region of Utah, Wyoming, Idaho, Oregon, Nevada, and California, is facing an uncertain water future due to climate change. The northern half of the Great Basin, which includes northern Utah, is located in the center of the El Niño–Southern Oscillation (ENSO) dipole. ENSO is a well-known climatic teleconnection between sea surface temperatures and the atmosphere in the equatorial Pacific Ocean which affects global weather patterns (Troup, 1965; Horel and Wallace, 1981). The occurrence of precipitation in the Great Basin in any given year is dependent on both the phase of ENSO and the phase of the Pacific Decadal Oscillation (PDO), as the phase of the PDO shifts the ENSO dipole either north or south (Wise, 2010; Brown, 2011). Due to its complex terrain, the majority of the water used by those who live in the region is dependent on the snowpack that is stored in the mountains and released throughout the year via the reservoir system. This semi-arid region is already experiencing inconsistent water availability throughout any given year due to the drastically different number of winter precipitation events from year to year. The ability to statistically model the occurrence of precipitation and air temperature is imperative to better forecast potential changes in future water availability as the climate changes. In this study, we introduce a stochastic harmonic autoregressive parametric (SHArP) weather generator, which statistically models meteorological variables (in this case, the occurrence and amount of precipitation and maximum air temperature). The model can be used to investigate how the future of the Great Basin may be impacted by climate change and to understand the meteorological extremes that are likely to play a part in that impact.

While the outputs of both statistically-based stochastic weather generators (SWGs) and dynamically-based global climate models (GCMs) are used in climate impacts studies, there are major differences between them. SWGs work on a point-scale, or on a point-scale expanded via multisite generalization to a basin-scale,

whereas GCMs work on a broad regional scale and can be downscaled to the basin or smaller scale. GCMs have difficulty capturing detail in areas of complex terrain, including the Great Basin, which is characterized by its basin-and-range topography (e.g., Thompson and Burke, 1974). SWGs also have a faster computational time than GCMs, which can take upwards of months to complete a single run. GCMs are very computationally expensive compared to SWGs, and thus, there are not many GCM runs available for analysis. GCMs also have difficulty capturing the very low-frequency (century-scale) connections between the Pacific Ocean and the Great Basin. The performance of state-of-the-art GCMs has been evaluated in terms of ability to capture the "extremes" in precipitation and temperature, and it has been found that GCMs poorly capture the extremes, though they perform better at temperature extremes than precipitation (Kiktev et al., 2007).

SWGs alleviate some limitations of GCMs and were introduced as a way to overcome a lack of observational meteorological data and problems associated with missing data both temporally and spatially (Wilks and Wilby, 1999; Wilks, 2008). In addition, they have been used to better understand the uncertainties associated with future climate (e.g., Wilks, 1992; Forsythe et al., 2014). These statistical models generate synthetic time series of precipitation and in some cases also air temperature and solar radiation, which statistically resemble the data used to force the model – usually daily observational weather data (Wilks and Wilby, 1999). There have been a multitude of early studies on SWGs that solely generate precipitation occurrence and amount because air temperature and other meteorological variables are affected by whether precipitation occurred.

The first studies using stochastic simulators of weather data employed two-state, first-order Markov chain frameworks regarding precipitation (Bailey, 1964; Richardson, 1981; Roldàn and Woolhiser, 1982), meaning that the probability of precipitation occurrence on a given day is only dependent on whether precipitation occurred on the previous day. Precipitation amount was modeled separately, and maximum/minimum temperatures and solar radiation were modeled as a function of precipitation occurrence. Other studies involving SWGs considered a two-state, second-order Markov chain process (Stern and Coe, 1984; Wilks, 1999a).

Markov chains of higher order have been found to better capture dry spells than first-order Markov chains, thus providing more accurate results for most areas of the western U.S. where dry spells are common, such as the semi-arid Great Basin.

One limitation of the common SWGs is the ability to successfully capture non-stationary variability. Previous studies have found that over the western U.S., El Niño results in a wetter Southwest and a drier Northwest, while La Niña results in the opposite (Ropelewski and Halpert, 1986; Dettinger et al., 1998; Woolhiser, 2008). In addition, the Pacific Decadal Oscillation (PDO) also has significant impacts on precipitation in the western U.S. The PDO is linked to ENSO, which in turn affects how the different phases of ENSO will impact the western U.S. (Gershunov and Barnett, 1998; Gershunov et al., 1999; Mauget, 2003). Woolhiser (2008) introduced the idea of adding nonstationarity to the stochastic framework in order to capture the effects these major oceanic oscillations have on western U.S. precipitation. Essentially, perturbations given as time series of the oscillations were linearly added to the probability of precipitation, and the coefficients associated with each perturbation give information on the sensitivity of each of the oscillations (Woolhiser, 2008). We employ this method in this study and also include a trend to account for the changing climate.

In the SWG literature, simulation of daily maximum and minimum air temperature is usually conditioned on whether the day is wet or dry. The most widely used method for simulating temperature is the method used by Richardson (1981). This method involves generating the standardized residual time series of temperature (maximum and mininum temperature; the study also included solar radiation) and using the multivariate generation model as described by Matalas (1967). These standardized residuals are assumed normally distributed, and the coefficients in the generating model are matrices containing the cross-correlations and auto-correlations between the residuals (Matalas, 1967). After generating the synthetic residuals, the wet- or dry-state means and standard deviations that were initially removed are reintroduced to yield daily values of the variables. The means and standard deviations depend on whether the day was wet or dry; they are assumed to be cyclostationary and are determined by fitting harmonics of the

annual cycle to observations (Richardson, 1981).

In addition to the common parametric SWGs described thus far, including the SWGs introduced by Matalas (1967) and Richardson (1981), recent studies have employed nonparametric SWGs and generalized linear models (GLMs). These SWGs are data-driven and involve either kernel density estimation (e.g., Rajagopalan et al., 1997; Harrold et al., 2003) or resampling via *k*-nearest neighbor (k-NN) bootstrapping (e.g., Rajagopalan and Lall, 1999; Caraway et al., 2014). These models do not rely on the statistical relationships applied in the parametric SWGs. They offer an alternative to the standard linear models presented in the parametric SWGs, which are unable to capture the nonlinear relationships between meteorological variables. The use of GLMs in SWGs, first introduced by Stern and Coe (1984), has also been increasing in popularity because they can easily model discrete variables and variables with non-normal distributions (Furrer and Katz, 2007). In addition, GLMs are especially useful tools because of their ability to treat ENSO and other major oceanic modes of variability as continuous variables (e.g. Chandler, 2005). More details behind GLMs can be found in McCullagh and Nelder (1989).

A limitation of the widely-used Richardson model is that its mean and standard deviation switch abruptly between wet- and dry-state values prescribed in advance of the simulation, and temperature is not simulated directly but rather through its residuals. This method inaccurately captures what occurs in reality, which are instead smooth, autocorrelated transitions between wet- and dry-state values. In this study, we introduce the mathematics and present illustrative results for a stochastic harmonic autoregressive parametric (SHArP) weather generator that is based on the Richardson model but that simulates temperature values directly with a mean that makes autocorrelated transitions between wet-and dry-state temperature values. Because of this innovation, the method described here better captures the temperature transitions between days with different precipitation states, including following frontal passages.

2.3 Data and Study Area

We chose to illustrate the SHArP weather generator using observations from the Salt Lake City International Airport (KSLC), which is located in the Great Basin. Its precipitation depends largely on a combination of the state of El Niño-Southern Oscillation (ENSO) and the state of the Pacific Decadal Oscillation (PDO) (Wise, 2010; Brown, 2011). The precipitation and temperature data used to force SHArP are daily observational data recorded at KSLC (40.78°N, 111.97°W) from 1 January 1948 to 31 December 2010 via the Global Historical Climatology Network (GHCN-Daily) provided by the National Centers for Environmental Information (obtained from www.ncdc.noaa.gov). In addition, we obtained GHCN-Daily precipitation and temperature data for four climatologically similar surrounding sites to illustrate the autocorrelated transitions during frontal passages. The domain map (see Fig. 2.1) shows the location of KSLC in addition to the four surrounding sites: Boise Air Terminal (KBOI) and Pocatello Regional Airport (KPIH) in Idaho, Elko Regional Airport (KEKO) in Nevada, and Grand Junction Regional Airport (KGJT) in Colorado.

Future precipitation and temperature output used to force SHArP are daily 0.125° gridded BCCA (bias correction constructed analog) projections from the CCSM4 model, which was part of the Coupled Model Intercomparison Project Phase 5 (CMIP5) multi-model ensemble (Maurer et al., 2007; Reclamation, 2013). We use the high emissions scenario (RCP 8.5) data, and they span from 1 January 2006 to 31 December 2100. We use the data starting from 1 January 2011 following the end of the observational data.

A day was considered "wet" and given value $\chi = 1$ if the total precipitation on that day reached at least 0.25 mm (approximately 0.01 inches), corresponding to the minimum depth recorded by rain gauges. Otherwise, the day was considered dry and given value $\chi = 0$. The χ vector was determined from the precipitation time series, and this provided the precipitation occurrence needed to model temperature with SHArP. In this study, we use and generate only maximum surface air temperature at a single site. Generalization to multiple variables at multiple sites has been completed, and the formulation will be presented in a future manuscript.

2.4 Simulation of Maximum Air Temperature and Precipitation

The method introduced here is based on the Richardson (1981) method described in the Introduction. The Richardson method is a linear equation given by

$$\chi_{p,i}(j) = A\chi_{p,i-1} + B\epsilon_{p,i}(j),$$

where $\chi_{p,i}(j)$ is a 3×1 matrix containing the residuals for day i of year p and $\chi_{p,i-1}(j)$ is a 3×1 matrix containing the residuals for day $i-1$ of year p; j refers to the variable of interest (Richardson simulated three: maximum temperature, minimum temperature, and solar radiation). A and B are 3×3 matrices that contain the correct serial and cross-correlation coefficients. The mean and standard deviation of the variables are removed, and the residuals are simulated. The model makes abrupt switches between wet- and dry-state values because of the prescribed means and standard deviations prior to simulation, which are then used to determine the true values after simulation.

SHArP is based on the observation shown below that temperature makes autocorrelated transitions between wet- and dry-state means with characteristic annual cycles, while subject to random fluctuations associated with frontal passages. For maximum temperature at a single site, the linear model is

$$T_{k+1} = aT_k + b_k + c_k\epsilon_k, \tag{2.1}$$

where a is a coefficient assumed constant, and b_k and c_k are coefficients that depend on day k. The b_k coefficient captures the mean, annual cycle, and trend. Errors ϵ_k are independent and identically distributed (i.i.d.) random standard normals. The temperature on day $k+1$ is dependent on the temperature on day k, where k ranges from 0 to $K-1$ (K being the length of the simulation). We begin the simulations by taking the first temperature value from the training data as T_0, but this could also be drawn from an appropriate distribution.

2.4.1 Maximum Likelihood Estimation

We begin with a simplified case where c does not depend on k. We assume the temperature entries T_1 to T_K are multivariate normals, and the joint density

function is given by

$$f(T_1, ..., T_K) = \frac{1}{(2\pi)^{K/2} c} \exp \frac{-(DT - B)'(DT - B)}{2c^2}, \tag{2.2}$$

where D is the $K \times K$ matrix:

$$D = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -a & 1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -a & 1 \end{bmatrix}$$

and B and T are the $K \times 1$ vectors:

$$B = \begin{bmatrix} aT_0 + b_0 \\ b_1 \\ \vdots \\ b_{K-2} \\ b_{K-1} \end{bmatrix}, T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_{K-1} \\ T_K \end{bmatrix}.$$

The mean is given by $D^{-1}B$, and the dry and wet day means are shown with their corresponding composite annual cycles from the KSLC training data in Fig. 2.2. Note the higher variability associated with wet days versus dry days. To restrict the model to a reasonable number of parameters, we give structure to the b_k values by giving them a trend and harmonics

$$\begin{aligned} b_k = \gamma_{\chi_{k+1}} + \alpha k + \beta_{\chi_{k+1}} \cos(2\pi k/\tau) + \beta'_{\chi_{k+1}} \sin(2\pi k/\tau) \\ + \delta_{\chi_{k+1}} \cos(4\pi k/\tau) + \delta'_{\chi_{k+1}} \sin(4\pi k/\tau), \end{aligned} \tag{2.3}$$

where τ is the period, assumed to be 365 days. We include two harmonics to illustrate how b_k can be generalized to include any number of harmonics. A log-likelihood ratio test can be performed to determine statistical significance of additional harmonics.

We applied a maximum likelihood estimate (MLE) to the joint density function, which involves maximizing (2.2) or minimizing its negative log

$$c^{-2}(DT - B)'(DT - B) + 2K \log c. \tag{2.4}$$

We first minimize $(DT - B)'(DT - B)$ to get the MLEs for the D and B matrices. This returns the sum of squared errors

$$(DT - B)'(DT - B) = \sum_{k=0}^{K-1} (aT_k + b_k - T_{k+1})^2, \tag{2.5}$$

where b_k is given in (2.3).

Taking derivatives in (2.5) with respect to a and each of the parameters in b_k and setting them equal to zero gives the following twelve equations:

$$\begin{aligned}
&\sum_{k=0}^{K-1} T_k(aT_k + b_k - T_{k+1}) = 0,\\
&\sum_{k=0}^{K-1} k(aT_k + b_k - T_{k+1}) = 0,\\
&\sum_{k=0}^{K-1} (aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 0\} = 0,\\
&\sum_{k=0}^{K-1} (aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 1\} = 0,\\
&\sum_{k=0}^{K-1} \cos(2\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 0\} = 0,\\
&\sum_{k=0}^{K-1} \cos(2\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 1\} = 0,\\
&\sum_{k=0}^{K-1} \sin(2\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 0\} = 0,\\
&\sum_{k=0}^{K-1} \sin(2\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 1\} = 0,\\
&\sum_{k=0}^{K-1} \cos(4\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 0\} = 0,\\
&\sum_{k=0}^{K-1} \cos(4\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 1\} = 0,\\
&\sum_{k=0}^{K-1} \sin(4\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 0\} = 0,
\end{aligned}$$

$$\sum_{k=0}^{K-1} \sin(4\pi k/\tau)(aT_k + b_k - T_{k+1})\mathbb{1}\{\chi_{k+1} = 1\} = 0,$$

where $\mathbb{1}$ is an indicator function which takes the value of one if the condition in brackets is met and zero otherwise. This is a linear system of 12 equations and 12 unknowns, which we solve numerically.

We then minimize (2.4) as a function of c. Taking a derivative in c yields

$$-2c^{-3}(DT - B)'(DT - B) + 2Kc^{-1}. \tag{2.6}$$

The derivative has a unique point at which it vanishes:

$$c = \sqrt{K^{-1}(DT - B)'(DT - B)}, \tag{2.7}$$

which is both the MLE value and least squares estimation (LSE) value. However, constant c tends to overestimate the variance in the summer and underestimate it in the winter (see panels b and c of Fig. 2.3), motivating a seasonally-varying c denoted by c_k, as in (2.1). The seasonally-varying c_k makes the MLE nonlinear in the parameters, so we proceed by taking an LSE approach where linear analytical expressions can be obtained.

2.4.2 Least Squares Estimation with Varying c_k

When c_k does not depend on k, the LSE for the parameters in b_k and a is equivalent to the MLE and system of 12 equations in Section 2.4.1. Now, we assume that c_k^2 has a cyclostationary structure similar to b_k but without a trend. Its formulation is given by

$$\begin{aligned} c_{k,0}^2 = \rho_0 &+ \epsilon_0 \cos(2\pi k/\tau) + \epsilon_0' \sin(2\pi k/\tau) \\ &+ \kappa_0 \cos(4\pi k/\tau) + \kappa_0' \sin(4\pi k/\tau) \end{aligned} \tag{2.8}$$

for dry days and

$$\begin{aligned} c_{k,1}^2 = \rho_1 &+ \epsilon_1 \cos(2\pi k/\tau) + \epsilon_1' \sin(2\pi k/\tau) \\ &+ \kappa_1 \cos(4\pi k/\tau) + \kappa_1' \sin(4\pi k/\tau) \end{aligned} \tag{2.9}$$

for wet days. Here, k also varies from 0 to $K - 1$. However, because we assume that c_k is cyclostationary with no trend, it is sufficient to specify $c_{k,0}$ and $c_{k,1}$ only

for $k = 0, ..., \tau - 1$. Our strategy to estimate $c_{k,0}$ and $c_{k,1}$ is to align the data by day of year $j = 0, ..., \tau - 1$ and segregate it according to the precipitation sequence. This yields the MLE (and LSE) estimators

$$\hat{c}_{j,0} = \sqrt{N_{j,0}^{-1}(DT - B)'_{j,0}(DT - B)_{j,0}} \quad \text{and} \tag{2.10}$$
$$\hat{c}_{j,1} = \sqrt{N_{j,1}^{-1}(DT - B)'_{j,1}(DT - B)_{j,1}},$$

where $(DT - B)_{j,0}$ is the $K/\tau \times 1$ vector populated with $(DT - B)_k$ if $\chi_{k+1} = 0$ and with zero if $\chi_{k+1} = 1$, where $k = j, j + \tau, j + 2\tau, ..., j + K - \tau$. Similarly, $(DT - B)_{j,1}$ is the $K/\tau \times 1$ vector populated with $(DT - B)_k$ if $\chi_{k+1} = 1$ and with zero if $\chi_{k+1} = 0$, where $k = j, j + \tau, j + 2\tau, ..., j + K - \tau$. $N_{j,0}$ is the number of times $\chi_{k+1} = 0$, and $N_{j,1}$ is the number of times $\chi_{k+1} = 1$.

Once we have estimated $c_{j,0}$ and $c_{j,1}$, we use the LSE method to estimate the parameters in equations (2.8) and (2.9). Specifically, we minimize

$$\sum_{j=0}^{\tau-1}(\rho_0 + \epsilon_0 \cos(2\pi j/\tau) + \epsilon'_0 \sin(2\pi j/\tau) \tag{2.11}$$
$$+\kappa_0 \cos(4\pi j/\tau) + \kappa'_0 \sin(4\pi j/\tau) - \hat{c}_{j,0}^2)^2$$

and

$$\sum_{j=0}^{\tau-1}(\rho_1 + \epsilon_1 \cos(2\pi j/\tau) + \epsilon'_1 \sin(2\pi j/\tau) \tag{2.12}$$
$$+\kappa_1 \cos(4\pi j/\tau) + \kappa'_1 \sin(4\pi j/\tau) - \hat{c}_{j,1}^2)^2.$$

Taking derivatives in each of the parameters in (2.11) and (2.12) and setting them equal to zero yields equations that are familiar from Fourier analysis. For dry days, we have

$$\hat{\rho}_0 = \tau^{-1} \sum_{j=0}^{\tau-1} \hat{c}_{j,0}^2,$$
$$\hat{\epsilon}_0 = \frac{2}{\tau} \sum_{j=0}^{\tau-1} \hat{c}_{j,0}^2 \cos(2\pi j/\tau),$$
$$\hat{\epsilon}'_0 = \frac{2}{\tau} \sum_{j=0}^{\tau-1} \hat{c}_{j,0}^2 \sin(2\pi j/\tau),$$

$$\hat{\kappa}_0 = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,0}^2\cos(4\pi j/\tau),$$

$$\hat{\kappa}_0' = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,0}^2\sin(4\pi j/\tau),$$

and for wet days, we have

$$\hat{\rho}_1 = \tau^{-1}\sum_{j=0}^{\tau-1}\hat{c}_{j,1}^2,$$

$$\hat{\epsilon}_1 = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,1}^2\cos(2\pi j/\tau),$$

$$\hat{\epsilon}_1' = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,1}^2\sin(2\pi j/\tau),$$

$$\hat{\kappa}_1 = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,1}^2\cos(4\pi j/\tau),$$

$$\hat{\kappa}_1' = \frac{2}{\tau}\sum_{j=0}^{\tau-1}\hat{c}_{j,1}^2\sin(4\pi j/\tau).$$

The parameters are inserted back into (2.8) or (2.9) to generate the synthetic temperature series using the linear model (2.1). Example simulations with the seasonally-varying c_k are shown in the right column of Fig. 2.3. Note how seasonally-varying c_k better captures the low variability in the summer and high variability in the winter. The dry and wet c_k curves are shown with composite annual cycles of the standard deviation of the noise in Fig. 2.4. These curves highlight the larger variability associated with wet days as well as the larger variability associated with the transition seasons (spring and fall) featuring strong frontal temperature contrasts.

2.4.3 Simulation of Precipitation

In this section, for completeness, we provide formulation for simulation of daily precipitation in a manner compatible with the temperature model introduced above. Our formulation largely follows Woolhiser (2008), except here we allow for trends in the Markov chain parameters. The probability of precipitation occurrence is determined with a two-state (wet or dry), second-order Markov chain,

which means that the probability of precipitation on a given day depends on the precipitation state on the previous two days as follows

$$p_{ij0}(t) = P\left\{\chi_t = 0 | \chi_{t-1} = j,\ \chi_{t-2} = i\right\};\ \ t = 1, 2, \ldots, 365M, \tag{2.13}$$

where M is the number of years. If we assume cyclostationarity, then the p_{ij0} terms are periodic, meaning $p_{ij0}(t + K365) = p_{ij0}(t)$ for any integer K. To account for nonstationarity associated with low-frequency oceanic forcing plus any trend, we define perturbed versions of (2.13)

$$\tilde{p}'_{ij0}(t) = \tilde{p}_{ij0}(t) + b_0^{(ij0)} + b_1^{(ij0)} t + b_2^{(ij0)} S_1(t - \tau_1) + b_3^{(ij0)} S_2(t - \tau_2)) \tag{2.14}$$

where $\{b_0, b_1\}$ enable a trend, S_1 and S_2 are oceanic forcing with periodicity of 3-7 years (ENSO) and 10-15 years (PDO), respectively, and the τ terms are positive lags (i.e., variations in ENSO and PDO indices may lead their effects on precipitation by τ months).

2.5 Comparison to the Richardson Method

Because the Richardson method of simulating stochastic temperature (referred to as the multivariate generation model) is the most widely-used parametric method in the field and the one upon which SHArP builds, it is a useful point of comparison. The Richardson method is essentially an autoregressive process that simulates standardized residuals; the details of this method can be found in Richardson (1981) and Matalas (1967). The Richardson method prescribes the means and standard deviations of the data (for wet and dry days) prior to simulation via a harmonic fit and then reintroduces them after simulating standardized residuals. This causes the model mean and standard deviation to abruptly switch between wet- and dry-state values. The model we introduce here (2.1) also has wet- and dry-state harmonics (b_k) and noise amplitudes (c_k) prescribed in advance, but the mean of the model ($D^{-1}B$) and standard deviation make autocorrelated, and hence more realistic, transitions via the parameter a in D.

We highlight the difference between the methods in Fig. 2.5, which compares the composite synthetic temperature simulated by the two models to the obser-

vational temperature for precipitation occurrence sequences of dry-dry-wet-wet-dry-dry for each season. The observational temperature reflects a typical cold frontal passage in each season (e.g., Shafer and Steenburgh, 2008). In general, the observed maximum temperature increases shortly before the frontal passage due to southerly flow and warm air advection; on the first day of precipitation, the maximum temperature decreases modestly. On the second day of precipitation, the temperature continues to decrease, and it slowly rebounds following the precipitation event. SHArP is able to capture this overall pattern. In contrast, the abrupt switching between wet- and dry-state means in the Richardson model results in an unrealistically large decrease in temperature on the first day of precipitation, followed by minimal change on the second day (actually zero change with large enough sample). While there is little to no seasonal bias in the Richardson model, there is a bias in the temperature around frontal passages. The temperature bias in the Richardson model is up to 4°C following frontal passages, and SHArP is able to reduce that bias by two degrees in three seasons.

Although the Richardson framework as originally formulated does not contain a trend term, one could be added in principle. One approach would be to fit the trend by LSE and remove it prior to estimating the annual cycles of the mean and residual standard deviations, and then adding the trend back in after generating the simulated temperatures. In contrast to this multistep approach involving removing components, fitting, simulating, and reintroducing components, the model presented here involves only fitting and simulation because all variations are captured in the fit formulation, including a trend term which is incorporated into b_k. Trended output from observations (1948-2010) and future BCCA CCSM4 high emissions scenario output (2011-2100) is shown in Fig. 2.6a with an example corresponding realization from the temperature model presented here shown in Fig. 2.6b.

2.6 Discussion and Conclusions

This study presents a new linear model for simulating stochastic temperature realizations called SHArP, and the method was illustrated for maximum temper-

ature at a single site within the Great Basin. We first considered a simplified version of the model with a constant noise coefficient, c, and applied MLE to obtain its parameters. However, this constant c compromised between the variance in the summer and the variance in the winter, which resulted in a simulation that did not adequately capture the seasonal variance found in the observations. A seasonally-varying noise coefficient, c_k, rendered the MLE nonlinear, and we presented analytical solutions via LSE. The resulting temperature realization more closely matched that of observations, with increased wintertime variance and decreased summertime variance.

Further realism may also be possible by relaxing assumptions used here. For example, we assume the amplitude of noise, c_k, to be annually cyclostationary but without trend. It is possible for the noise to have similar nonstationarity due to ENSO and PDO. Curvilinearity (a trend) and variables related to ENSO and PDO could be added to the c_k^2 equations (if the area of interest is in a region where these oceanic modes play a major role) and solved using the same LSE method. A nonlinear trend could also be added to αk portion of the b_k equation, making it $\alpha_1 k + \alpha_2 k^2$. We also assume that temperature depends only on itself and precipitation occurrence, but precipitation amount and climate teleconnections that influence air mass trajectories may be additionally important.

Even though this study is focused on only maximum temperature at a single site, we have generalized the method described to include minimum temperature in addition to maximum temperature at multiple sites. The linear model remains the same, but the scalar computations become matrix computations. We extended ideas described in Wilks (1998) and Wilks (1999b), where the sites themselves have spatial correlation but are generated independently of each other, by introducing spatial correlations in the c_k matrices but not in the a matrix. However, this method introduced an increased number of parameters in the variance-covariance matrix that required a nontrivial technique to mitigate the issue, and this will be described in a future manuscript.

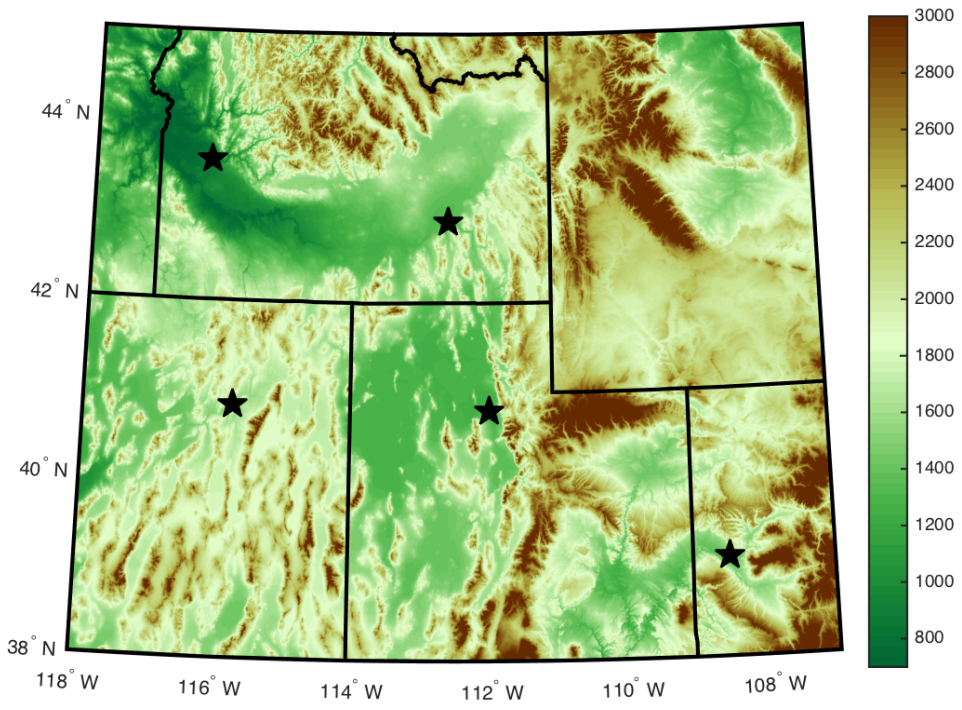

Figure 2.1. The study area: the eastern half of the Great Basin (which includes northern and western Utah, extreme southwestern Wyoming, extreme southern Idaho, and Nevada) and surrounding area. The stars indicate the location of the Salt Lake City International Airport (KSLC) and surrounding sites: Boise Air Terminal (KBOI) and Pocatello Regional Airport (KPIH) in Idaho, Elko Regional Airport (KEKO) in Nevada, and Grand Junction Regional Airport (KGJT) in Colorado. The colorbar indicates elevation in meters above sea level.

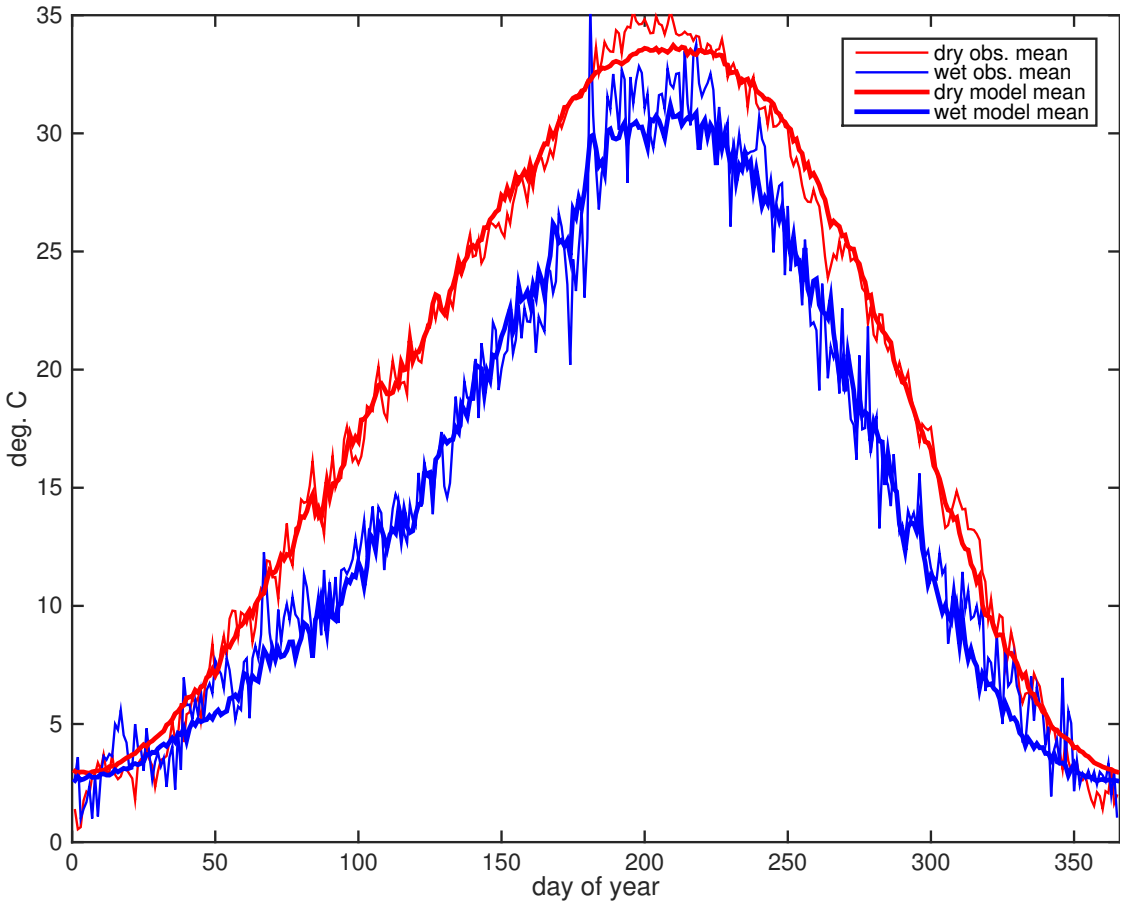

Figure 2.2. Annual composite of the observational and model means for dry days (red) and wet days (blue). Results are based on KSLC observations for years 1948 to 2010.

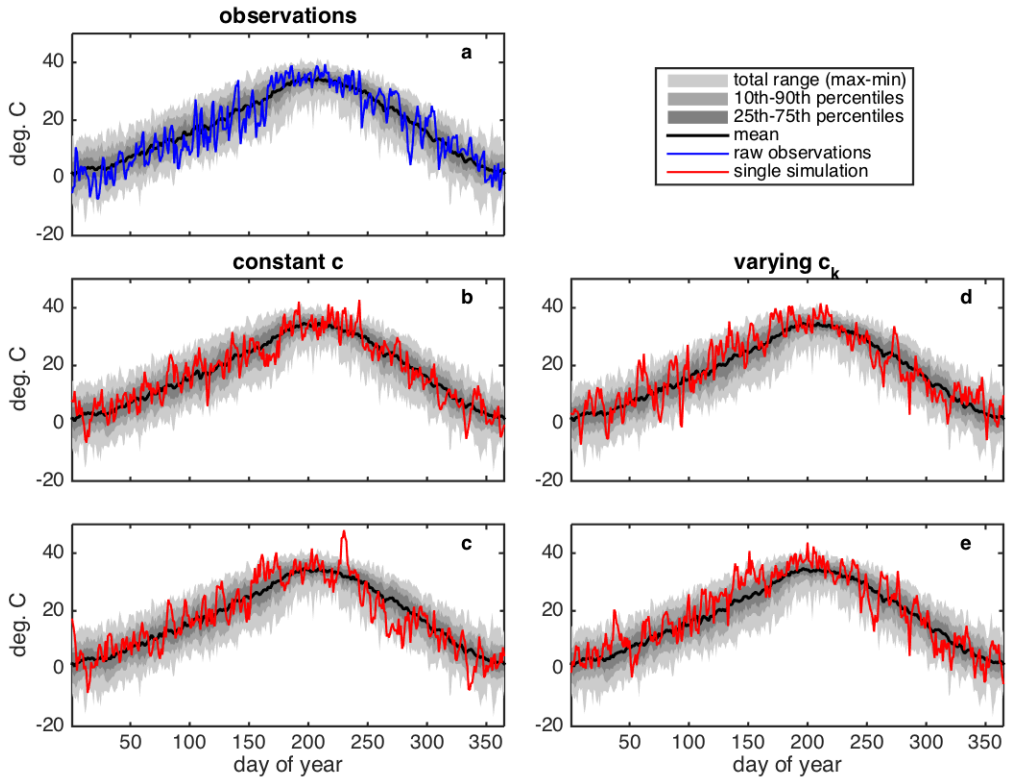

Figure 2.3. Illustration of the SHArP weather generator with (a) input observational data for comparison. The blue curve shows 2008 as an example year, and shading in each panel corresponds to percentiles of the historical data for 1948-2010. Two simulations of the temperature model with constant c are shown in (b) and (c), and two simulations of the temperature model with seasonally-varying c_k are shown in (d) and (e).

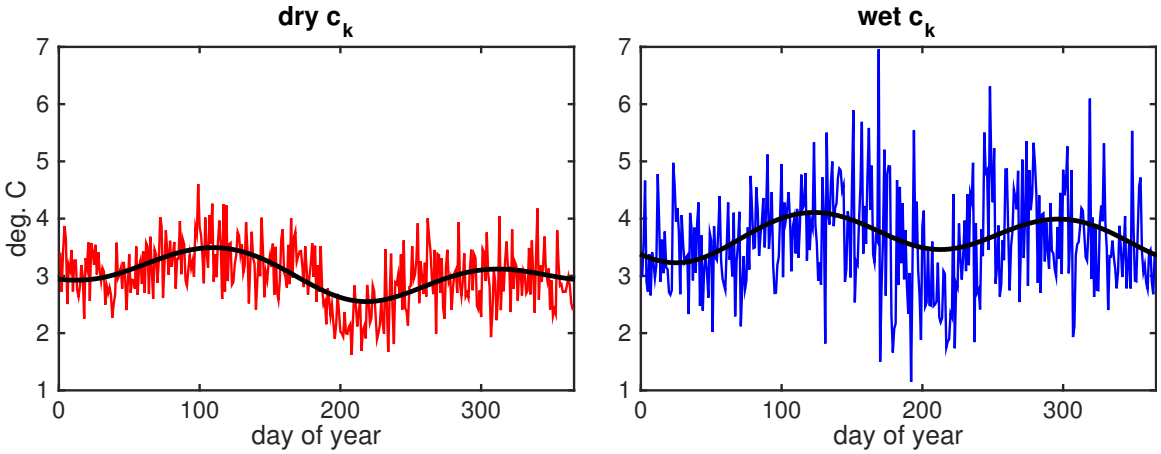

Figure 2.4. Seasonally-varying c_k curves for dry days and wet days (black lines) and standard deviations of the noise (colored lines). Note the relatively higher variability in the transitional seasons and overall higher variability associated with the wet days.

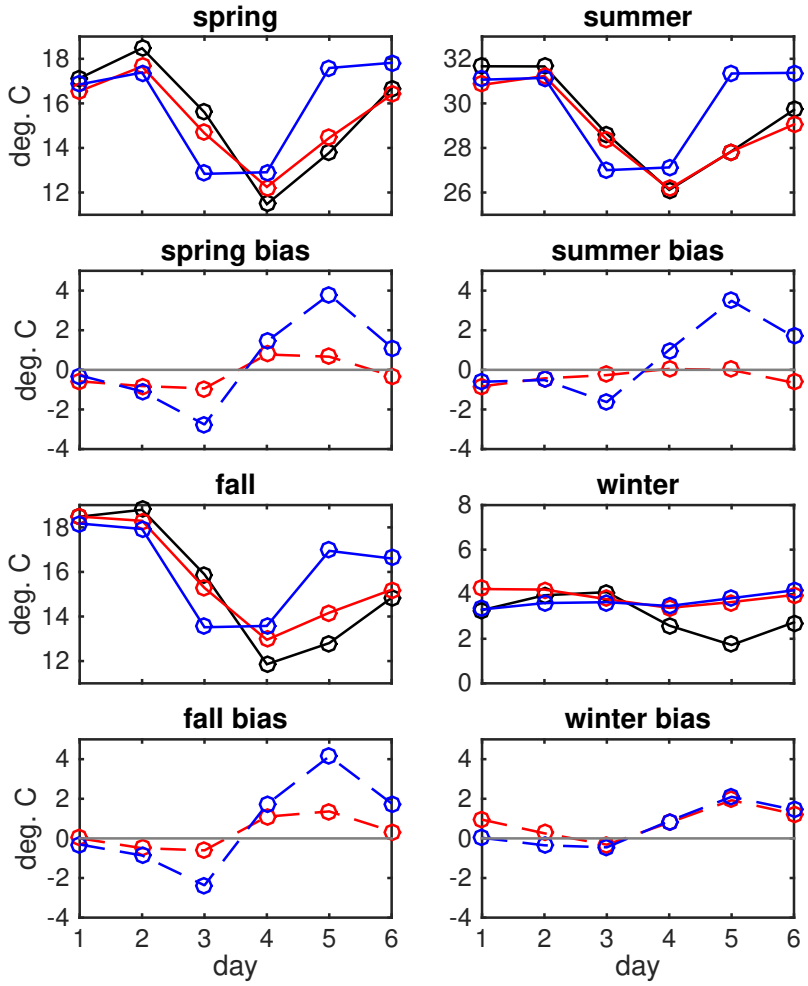

Figure 2.5. Composite observational temperature (black lines) and composite synthetic temperature for sets of days that follow the precipitation occurrence sequence dry-dry-wet-wet-dry-dry in each season. In addition, the bias for each season is shown immediately below. Composite is of each occurrence of this sequence at five climatologically-similar sites (see Fig. 2.1). The red lines indicate SHArP, the model presented here, and the blue lines indicate the Richardson model. The number of samples in each set is approximately 500.

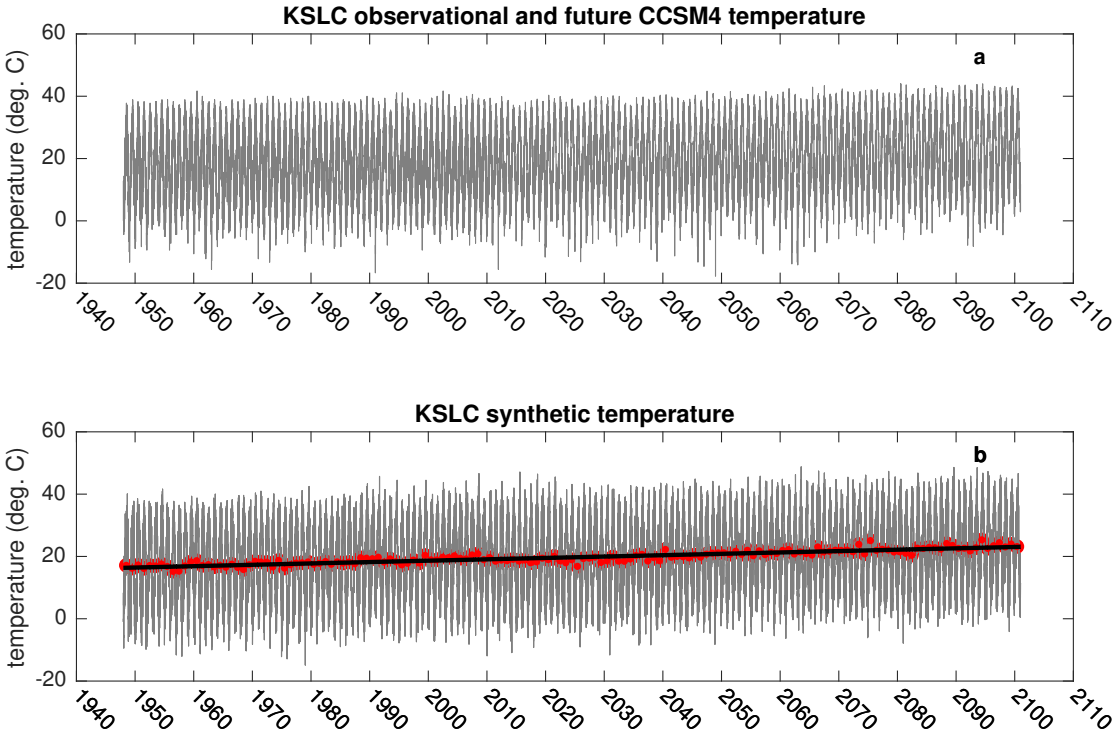

Figure 2.6. (a) KSLC observational GHCN-Daily maximum temperature (1948-2010) and BCCA CCSM4 high emissions (RCP8.5) maximum temperature output (2011-2100). (b) An example of trended stochastic maximum temperature simulated from 1948 to 2100 for KSLC. The simulation was trained on the data shown in the top panel. The red dots indicate the average annual maximum temperature for each year of the simulation.

2.7 References

Bailey, N. T. J., 1964: *The Elements of Stochastic Processes*. John Wiley, New York, 39 pp.

Brown, D. P., 2011: Winter circulation anomalies in the western United States associated with antecedent and decadal ENSO variability. *Earth Interactions*, **15 (3)**, 1–12, doi:10.1175/2010EI334.1, URL http://dx.doi.org/10.1175/2010EI334.1.

Caraway, N. M., J. L. McCreight, and B. Rajagopalan, 2014: Multisite stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *Journal of Hydrology*, **508**, 197 – 213, doi:http://dx.doi.org/10.1016/j.jhydrol.2013.10.054, URL http://www.sciencedirect.com/science/article/pii/S0022169413007981.

Chandler, R. E., 2005: On the use of generalized linear models for interpreting climate variability. *Environmetrics*, **16 (7)**, 699–715, doi:10.1002/env.731, URL http://dx.doi.org/10.1002/env.731.

Dettinger, M. D., D. R. Cayan, H. F. Diaz, and D. M. Meko, 1998: North–south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate*, **11 (12)**, 3095–3111, doi:10.1175/1520-0442(1998)011⟨3095:NSPPIW⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0442(1998)011<3095:NSPPIW>2.0.CO;2.

Forsythe, N., H. Fowler, S. Blenkinsop, A. Burton, C. Kilsby, D. Archer, C. Harpham, and M. Hashmi, 2014: Application of a stochastic weather generator to assess climate change impacts in a semi-arid climate: The Upper Indus Basin. *Journal of Hydrology*, **517**, 1019 – 1034, doi:10.1016/j.jhydrol.2014.06.031, URL http://www.sciencedirect.com/science/article/pii/S0022169414004910.

Furrer, E. M. and R. W. Katz, 2007: Generalized linear modeling approach to stochastic weather generators. *Climate Research*, **34 (2)**, 129–144, URL http://www.int-res.com/abstracts/cr/v34/n2/p129-144/.

Gershunov, A. and T. P. Barnett, 1998: Interdecadal modulation of ENSO teleconnections. *Bulletin of the American Meteorological Society*, **79 (12)**, 2715–2725, doi:10.1175/1520-0477(1998)079⟨2715:IMOET⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0477(1998)079<2715:IMOET>2.0.CO;2.

Gershunov, A., T. P. Barnett, and D. R. Cayan, 1999: North Pacific interdecadal oscillation seen as factor in ENSO-related North American climate anomalies. *Eos, Transactions American Geophysical Union*, **80 (3)**, 25–30, doi:10.1029/99EO00019, URL http://dx.doi.org/10.1029/99EO00019.

Harrold, T. I., A. Sharma, and S. J. Sheather, 2003: A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, **39 (12)**, n/a–n/a, doi:10.1029/2003WR002570, URL http://dx.doi.org/10.1029/2003WR002570, 1343.

Horel, J. D. and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Monthly Weather Review*, **109 (4)**, 813–829, doi:10.1175/1520-0493(1981)109⟨0813:PSAPAW⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;2.

Kiktev, D., J. Caesar, L. V. Alexander, H. Shiogama, and M. Collier, 2007: Comparison of observed and multimodeled trends in annual extremes of temperature and precipitation. *Geophysical Research Letters*, **34 (10)**, n/a–n/a, doi:10.1029/2007GL029539, URL http://dx.doi.org/10.1029/2007GL029539.

Matalas, N. C., 1967: Mathematical assessment of synthetic hydrology. *Water Resources Research*, **3 (4)**, 937–945, doi:10.1029/WR003i004p00937, URL http://dx.doi.org/10.1029/WR003i004p00937.

Mauget, S. A., 2003: Intra- to multidecadal climate variability over the continental United States: 1932–99. *Journal of Climate*, **16 (13)**, 2215–2231, doi:10.1175/2751.1, URL http://dx.doi.org/10.1175/2751.1.

Maurer, E. P., L. Brekke, T. Pruitt, and P. B. Duffy, 2007: Fine-resolution climate projections enhance regional climate change impact studies. *Eos, Transactions American Geophysical Union*, **88 (47)**, 504–504, doi:10.1029/2007EO470006, URL http://dx.doi.org/10.1029/2007EO470006.

McCullagh, P. and J. Nelder, 1989: *Generalized Linear Models*. 2nd ed., Chapman & Hall, London.

Rajagopalan, B. and U. Lall, 1999: A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, **35 (10)**, 3089–3101, doi:10.1029/1999WR900028, URL http://dx.doi.org/10.1029/1999WR900028.

Rajagopalan, B., U. Lall, and D. G. Tarboton, 1997: Evaluation of kernel density estimation methods for daily precipitation resampling. *Stochastic Hydrology and Hydraulics*, **11 (6)**, 523–547, doi:10.1007/BF02428432, URL http://dx.doi.org/10.1007/BF02428432.

Reclamation, 2013: Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of downscaled CMIP5 climate projections, comparion with preceding information, and summary of user needs. Tech. rep., U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, Colorado. 47pp.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17 (1)**, 182–190, doi:10.1029/WR017i001p00182, URL http://dx.doi.org/10.1029/WR017i001p00182.

Roldàn, J. and D. A. Woolhiser, 1982: Stochastic daily precipitation models: 1. A comparison of occurrence processes. *Water Resources Research*, **18 (5)**, 1451–1459.

Ropelewski, C. F. and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review*, **114 (12)**, 2352–2362, doi:10.1175/1520-

0493(1986)114⟨2352:NAPATP⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

Shafer, J. C. and W. J. Steenburgh, 2008: Climatology of strong intermountain cold fronts. *Monthly Weather Review*, **136 (3)**, 784–807, doi:10.1175/2007MWR2136.1, URL http://dx.doi.org/10.1175/2007MWR2136.1.

Stern, R. D. and R. Coe, 1984: A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, **147 (1)**, 1–34, URL http://www.jstor.org/stable/2981736.

Thompson, G. A. and D. B. Burke, 1974: Regional geophysics of the basin and range province. *Annual Review of Earth and Planetary Sciences*, **2**, 213–238.

Troup, A. J., 1965: The 'southern oscillation'. *Quarterly Journal of the Royal Meteorological Society*, **91 (390)**, 490–506, doi:10.1002/qj.49709139009, URL http://dx.doi.org/10.1002/qj.49709139009.

Wilks, D., 1992: Adapting stochastic weather generation algorithms for climate change studies. *Climatic Change*, **22 (1)**, 67–84, doi:10.1007/BF00143344, URL http://dx.doi.org/10.1007/BF00143344.

Wilks, D., 1998: Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, **210 (1–4)**, 178 – 191, doi:10.1016/S0022-1694(98)00186-3, URL http://www.sciencedirect.com/science/article/pii/S0022169498001863.

Wilks, D., 1999a: Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, **93 (3)**, 153–169, URL http://www.sciencedirect.com/science/article/pii/S0168192398001257.

Wilks, D., 1999b: Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, **96 (1–3)**, 85 – 101, doi:10.1016/S0168-1923(99)00037-4, URL http://www.sciencedirect.com/science/article/pii/S0168192399000374.

Wilks, D. S., 2008: High-resolution spatial interpolation of weather generator parameters using local weighted regressions. *Agricultural and Forest Meteorology*, **148 (1)**, 111 – 120, doi:10.1016/j.agrformet.2007.09.005, URL http://www.sciencedirect.com/science/article/pii/S0168192307002511.

Wilks, D. S. and R. L. Wilby, 1999: The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, **23 (3)**, 329–357, doi:10.1177/030913339902300302, URL http://ppg.sagepub.com/content/23/3/329.abstract, http://ppg.sagepub.com/content/23/3/329.full.pdf+html.

Wise, E. K., 2010: Spatiotemporal variability of the precipitation dipole transition zone in the western United States. *Geophysical Research Letters*, **37 (7)**, doi:10.1029/2009GL042193, URL http://dx.doi.org/10.1029/2009GL042193.

Woolhiser, D. A., 2008: Combined effects of the Southern Oscillation Index and the Pacific Decadal Oscillation on a stochastic daily precipitation model. *Journal of Climate*, **21 (5)**, 1139–1152, doi:10.1175/2007JCLI1862.1, URL `http://dx.doi.org/10.1175/2007JCLI1862.1`.

CHAPTER 3

MULTISITE GENERALIZATION OF THE SHARP WEATHER GENERATOR

3.1 Abstract

Generalization of point-scale stochastic weather generators to simultaneously produce output at multiple sites provides more powerful support for hydrology and climate change impacts studies. Generalization preserves the statistical properties of each individual site while maintaining the spatial correlation over the domain. Here, the generalization of both the daily precipitation and temperature components of the stochastic harmonic autoregressive parametric (SHArP) weather generator is presented. The generalization process for temperature involves conversion of vector time series to matrix time series that capture between-site covariances of maximum and minimum daily temperature. Between-site temperature covariances depend on spatial precipitation occurrence patterns, of which there are 2^M for M sites. To dramatically reduce the number of covariance matrices that drive temperature, multisite SHArP uses empirical orthogonal function analysis to categorize the precipitation occurrence patterns, and harmonic smoothing to reduce the number of parameters describing the temporal evolution (annual cycle) of the elements in the covariance matrices. For precipitation simulation, existing methods are used, and a trend term is added to the occurrence and amount parameters. The multisite generalization of the framework is illustrated by simulating stochastic historical and future temperature and precipitation for a transect across complex terrain over northern Utah.

3.2 Introduction

Stochastic weather generators (SWGs) were primarily introduced to simulate daily meteorological variables, namely precipitation and temperature, that are sta-

tistically similar to the observed data at the location in question. SWGs are especially useful tools for hydrologists, climate scientists, agriculturalists, ecologists, planners, engineers, and practitioners in related fields given missing meteorological data or an interest in ensemble statistics (e.g., for uncertainty analysis). The development of SWGs often begins with the precipitation process since most other meteorological variables depend on whether or not precipitation occurred, and the addition of air temperature is a natural next step. SWGs are constructed to work on a point-scale, but in order to further capture variations between sites or examine hydrologic or climate change impacts on a broader scale, the methods need to be generalized to multiple sites. Generalization to multiple sites has its own set of challenges, especially as the number of sites increases.

Wilks (1998) introduced the widely known multisite generalization model of precipitation occurrence and amount based on chain-dependent processes (a two-state, second-order Markov chain for occurrence and a mixed exponential distribution for amount) that were described in Todorovic and Woolhiser (1975) and later applied in Richardson (1981). This is done by applying spatially correlated yet time-independent random vectors on the models of each individual site within the domain (Wilks, 1998). With this method, each site retains its own statistical properties while maintaining realistic correlations with the neighboring sites (Wilks, 1998).

Wilks (1999b) expanded on the multisite generalization method presented in Wilks (1998) by applying the method over an area with complex terrain in the western United States. In addition, it was expanded to include daily maximum and minimum temperature and solar radiation following Richardson (1981). Fitted correlation functions were used to capture the seasonal variations in this area, and this multisite generation was able to model the precipitation over complex terrain while preserving the spatial correlations found in nature (Wilks, 1999b). Later, Wilks (2009) showed the practicality of a spatially coherent SWG that interpolated parameters for single sites as described in Wilks (2008). In addition, the study was able to synchronize the gridded synthetic data to true weather data at reference stations within the domain and provide more realistic simulations for hydrologic

purposes.

Caraway et al. (2014) developed a nonparametric multisite stochastic weather generator using the k-nearest neighbor (K-NN) resampling approach. This model uses clustering of homogeneous sites in addition to Markov chain states to simulate precipitation at multiple sites within a heterogeneous watershed. While most present-day weather generators are parametric and based off of the work of Richardson (1981) and Wilks (1998), including SHArP (Smith et al., 2017) and MulGETS (Chen et al., 2014), the advantages of a nonparametric weather generator include the ability to capture the nonlinear variability that is missed in the linear parametric stochastic weather generators. Kleiber et al. (2012) introduced a generalized linear model (GLM) that uses spatial Gaussian processes to model the statistical parameters of precipitation over a domain. A similar nonparametric GLM for maximum and minimum temperature was also developed and is described in Kleiber et al. (2013).

In this study, we show the multisite generalization of the stochastic harmonic autoregressive parametric (SHArP) weather generator introduced in Smith et al. (2017). The mathematical formulation follows that of the single-site, single-temperature case in Smith et al. (2017), but there are major difference with the handling of the stochastic term and autocorrelation. In the single-site, single-temperature case, we used a temporally-varying noise coefficient that depended on whether the given day was wet or dry at the site. Having multiple sites introduces between-site covariances, which are found to depend on precipitation occurrence patterns whose number increases as 2^M for M sites. To circumvent this, we use empirical orthogonal function analysis to objectively categorize the precipitation patterns, yielding a compact set of matrices for driving between-site temperature covariance. We show how this approach results in reasonable temperature simulations from SHArP. An example application in the Western U.S. is used to illustrate the fidelity of the framework in complex terrain where precipitation patterns can change markedly over the study domain, and the simulation period is 1950-2099 to illustrate handling of trends.

3.3 Data and Study Area

The precipitation and temperature data used to force the SHArP weather generator are 0.125° bias corrected constructed analogs (BCCA) of daily CCSM4 output from the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Maurer et al., 2007; Reclamation, 2013). We used the historical output in this analysis, which spans from 1 January 1950 to 31 December 2005, as well as the future RCP 8.5 (high emissions scenario) output, which spans from 1 January 2006 to 31 December 2100. For illustration, we selected a transect of 30 sites from the west desert of Utah (40.8125°N, 113.6875°W) to the Uinta Mountains (40.8125°N, 110.0625°W) that runs through the point nearest the Salt Lake International Airport (KSLC; 40.8125°N, 111.9375°W; site 15 indicated by the star) (Fig. 3.1). Half of the sites are located in the "valley" while the other half of the sites are located in the mountains. These sites are located within the larger Great Basin, which is known for its semi-arid climate and basin-and-range topography (e.g., Thompson and Burke, 1974). We note that observations can of course be used to train SHArP as well, but we focus on downscaled model output to exercise and illustrate the framework's full spatial and temporal capabilities.

To account for the influence of oceanic modes on precipitation received in the Great Basin, we included two additional parameters in the precipitation portion of SHArP: one for the ENSO-like variability and one for the PDO-like variability. These data are bandpass-filtered, spatially-averaged historical CCSM4 sea surface temperature (SST) output, and we chose CCSM4 from the CMIP5 due to its skill in capturing oceanic influences on Great Basin precipitation as shown by Smith et al. (2015). These oceanic forcings affect precipitation occurrence directly, and hence indirectly affect temperature.

3.4 Multisite Simulation of Daily Maximum and Minimum Air Temperature

3.4.1 Model Formulation

The linear model for simulating multiple temperatures at multiple sites follows the SHArP linear model introduced in Smith et al. (2017) and is given by

$$\mathbf{T}_{k+1} = \mathbf{A}\mathbf{T}_k + \mathbf{B}_k + \mathbf{C}_k\boldsymbol{\epsilon}_k, \tag{3.1}$$

where $\mathbf{A}$ is a $2M \times 2M$ autocorrelation matrix for number of sites M, $\mathbf{B}_k$ is a $2M \times 1$ column vector that depends on day k, $\mathbf{C}_k$ is a $2M \times 2M$ positive definite matrix made up of noise coefficients, and $\boldsymbol{\epsilon}_k$ is a $2M \times 1$ column vector. Errors $\boldsymbol{\epsilon}_k$ are independent and identically distributed (i.i.d.) random vectors with entries that themselves are independent standard normals. The temperature on day $k+1$ is dependent on the temperature on day k, where k ranges from 0 to $K-1$ (K being the length of the simulation).

We assume that $\mathbf{A}$ is time-independent and block-diagonal

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{A}_M \end{bmatrix} \tag{3.2}$$

where $\mathbf{0}$ is a (2×2) zero matrix and the elements of

$$\mathbf{A}_m = \begin{bmatrix} a_{\text{max,max}} & a_{\text{max,min}} \\ a_{\text{min,max}} & a_{\text{min,min}} \end{bmatrix} \tag{3.3}$$

capture the dependence of maximum and minimum temperature at site m on the prior day's maximum and minimum temperature at site m, as indicated by the subscripts (e.g., $a_{\text{min,max}}$ is the dependence of minimum temperature on the preceding day's maximum temperature). With this structure for $\mathbf{A}$, between-site covariance is provided by $\mathbf{C}_k$.

We model the time dependence of each component of $\mathbf{B}_k$ using harmonics generally written as

$$\begin{aligned} b_k = \gamma_{\chi_{k+1}} + \alpha k + \beta_{\chi_{k+1}} \cos(2\pi k/\tau) + \beta'_{\chi_{k+1}} \sin(2\pi k/\tau) \\ + \delta_{\chi_{k+1}} \cos(4\pi k/\tau) + \delta'_{\chi_{k+1}} \sin(4\pi k/\tau), \end{aligned} \tag{3.4}$$

where τ is the period, assumed to be 365 days. Here, b_k is one of the $2M$ entries of $\mathbf{B}_k$. Coefficients γ, α, β, β', δ, and δ' are also entries of $2M \times 1$ vectors. The subscript χ_{k+1} indicates that b_k depends on whether day $k+1$ was wet ($\chi = 1$) or dry ($\chi = 0$).

As in the single-site SHArP, we apply a least squares estimation (LSE) to determine the parameters in $\mathbf{A}$ and $\mathbf{B}_k$. Because there are two temperatures (maximum and minimum) and $\mathbf{A}$ is now a matrix of four elements per site instead of a single coefficient, we minimize the residuals by differentiating with respect to 26 variables per site instead of the 12 in the single-temperature, single-site case. The $22M$ resulting equations related to the (precipitation occurrence spatial pattern) parameters of $\mathbf{B}_k$ from (3.4) are analogous to those presented in (Smith et al., 2017), and are omitted for brevity. The remaining $4M$ equations related to the four elements of $\mathbf{A}$ at each site follow the form

$$\sum_{k=0}^{K-1} T_{max,k}(a_{max,max}T_{max,k} + a_{max,min}T_{min,k} + b_{k,max} - T_{max,k+1}) = 0,$$
$$\sum_{k=0}^{K-1} T_{min,k}(a_{max,max}T_{max,k} + a_{max,min}T_{min,k} + b_{k,max} - T_{max,k+1}) = 0,$$
$$\sum_{k=0}^{K-1} T_{max,k}(a_{min,max}T_{max,k} + a_{min,min}T_{min,k} + b_{k,min} - T_{min,k+1}) = 0,$$
$$\sum_{k=0}^{K-1} T_{min,k}(a_{min,max}T_{max,k} + a_{min,min}T_{min,k} + b_{k,min} - T_{min,k+1}) = 0,$$

where $b_{k,max}$ and $b_{k,min}$ refer to the elements of $\mathbf{B}_k$ that correspond to T_{max} and T_{min}, respectively.

3.4.2 Specification of Parameters

In the single-site, single-temperature case, the noise coefficient $\mathbf{c}_k$ was a time-dependent vector that depended on whether the day was wet or dry. For multisite SHArP, $\mathbf{C}_k$ is a time-dependent, $2M \times 2M$ matrix that depends on the spatial pattern of precipitation occurrence. In the multisite case, the number of possible spatial patterns of precipitation is 2^M, which would yield an unmanageably large set of $\mathbf{C}_k$ matrices. We reduce this dramatically using empirical orthogonal function (EOF) analysis (e.g., Hannachi et al., 2007) of the precipitation occurrence spatial patterns (i.e., we calculate the eigenvectors of the $2M \times 2M$ spatial covariance matrix of occurrence). For the example here, we used the positive and negative polarity of the first two EOFs of occurrence to define four precipitation patterns (Fig. 3.2a,e), and then assigned each day to the pattern it most closely resem-

bled. To determine the closest match, the eigenvectors were quantized so that nonnegative components were assigned a value 1 and negative components were assigned a value 0 (Fig. 3.2b,f), and the Euclidian distance was calculated between the quantized eigenvector and the spatial pattern of precipitation occurrence on that day. For the study area here, the first quantized eigenvector captured all sites being wet in its positive polarity and all sites being dry in its negative polarity (Fig. 3.2b). The second quantized eigenvector captured the mountain sites being wet and the valley sites being dry in its positive polarity, and the reverse in its negative polarity (Fig. 3.2f). Example days assigned to the four patterns are shown in Figs. 3.2c,d and 3.2g,h.

We populate each of the four $\mathbf{C}_k$ matrices with the principal square root of the residuals $(\mathbf{T}_{k+1} - \mathbf{A}\mathbf{T}_k - \mathbf{B}_k)$ specific to the given day of year and one of the four precipitation occurrence patterns determined via the EOF analysis (there are as many $\mathbf{C}_k$ matrices as there are precipitation pattern EOFs). We then temporally smooth the entries of each $\mathbf{C}_k$ as in Smith et al. (2017) using Fourier analysis with the general equation

$$c_k = \rho + \epsilon \cos(2\pi k/\tau) + \epsilon' \sin(2\pi k/\tau) + \kappa \cos(4\pi k/\tau) + \kappa' \sin(4\pi k/\tau), \quad (3.5)$$

where c_k is a time-dependent element of one of the four $\mathbf{C}_k$ matrices. During estimation and simulation, each of the $\mathbf{C}_k$ matrices contains maximum and minimum temperature for each site, so are of size $2M \times 2M$ with row 1 corresponding to maximum temperature at site 1, row two corresponding to minimum temperature at site 1, row three corresponding to maximum temperature at site 2, and so on.

3.4.3 Illustrative Patterns and Simulations

Once we have determined the parameters in $\mathbf{A}$, $\mathbf{B}_k$, and $\mathbf{C}_k$, we simulate maximum and minimum temperature simultaneously at all sites. To illustrate the utility of the precipitation pattern-based matrices ($\mathbf{C}_k$), Fig. 3.3 contrasts the covariance structure of the stochastic residuals $(\mathbf{T}_{k+1} - \mathbf{A}\mathbf{T}_k - \mathbf{B}_k)$ between different precipitation patterns for selected seasons. For the stochastic residuals of maximum temperature during summer, variance is larger over the valley sites (1-15) than over the mountain sites (16-30) (Fig. 3.3a), and variance and covariance increase

almost uniformly in the transition from all-dry to all-wet (Fig. 3.3a,b). For the stochastic residuals of maximum temperature during winter, variance is larger over the mountain sites than over the valley sites (Fig. 3.3c), and variance and covariance increase more dramatically over the valley than over the mountains in the transition from all-dry to all-wet (Fig. 3.3c,d). For the stochastic residuals of minimum temperature during fall, variance peaks sharply at valley-mountain transition near site 19 (Fig. 3.3e). In the transition from all-dry to mountain wet / valley dry, the variance and covariance of the minimum temperature stochastic residuals increase almost uniformly (Fig. 3.3e,f), and the covariance of eastern mountain sites with sites to the west decreases sharply at the valley-mountain transition (Fig. 3.3f).

Turning attention from the stochastic residuals to the temperatures themselves, we now illustrate evolution of temperature during transitions between wet and dry conditions. As an example, we composite across sequences of three all-wet days followed by three all-dry days during July-September 1950-2100, and focus on four sites for visual clarity in the graphs (Fig. 3.4). In the training data and SHArP simulation, maximum and minimum temperature tend to progressively decrease with each additional wet day, and maximum temperature rebounds faster with the transition to dry conditions (Fig. 3.4a,c). The variance of maximum and minimum temperature is largest near the wet-to-dry transition, and its decrease with the transition to dry conditions is more pronounced for maximum temperature than for minimum temperature (Fig. 3.4b,d). Illustrating covariation on longer time scales for these same four sites, an annual cycle of maximum temperature is shown for a late-century year in Fig. 3.5a. SHArP, in addition to simulating a realistic annual cycle and variance at each site, provides realistic intersite covariation that is temporally synchronized by precipitation patterns (compare Fig. 3.5a,b). SHArP is also able to capture long-term trends with realistic intersite covariation as illustrated by annual mean minimum temperatures at the four sites (compare Figs. 3.5c,d).

3.5 Multisite Simulation of Daily Precipitation

3.5.1 Formulation and Parameter Estimation for Precipitation Occurrence

The precipitation model we use with SHArP largely follows formulations presented in Woolhiser (2008) and Wilks (2009), except we introduce a trend term in the perturbation of the Markov chain precipitation occurrence probabilities so that the framework can simulate climate change. We provide details leading up to the introduction of the trend here for completeness.

We model precipitation occurrence with a two-state (wet or dry), second-order Markov chain such that the probability of precipitation on any given day depends on the precipitation state on the previous two days:

$$p_{ij1}(t) = P\left\{\chi_t = 1 | \chi_{t-1} = j,\ \chi_{t-2} = i\right\};\ \ t = 1, 2, \ldots, 365Y, \tag{3.6}$$

where Y indicates the number of years. We use a second-order Markov chain as opposed to a first-order Markov chain because the former has been shown to better capture the occurrence of dry spells (e.g., Stern and Coe, 1984; Wilks, 1999a), which are common in the semi-arid region in this study. The unperturbed probability time series in (3.6) are cyclostationary, written as inverse logits, and found via maximum likelihood using a Newton-Raphson iterative procedure (Woolhiser, 2008). To illustrate the spatiotemporal patterns of these probability functions, the p_{011} values for each site over any given year are shown in Fig. 3.6a. The probability of precipitation is overall higher in the mountains than in the valleys, and the maximum in p_{011} for most sites occurs near day of year 100. The marked increase in p_{011} at the valley-to-mountain transition near site 15 motivates use of EOF analysis to categorize the precipitation occurrence patterns, and contributes to mountain versus valley contrast captured by EOF 2 (recall Figs. 3.2b,d).

We generalize the precipitation occurrence process to $m = 1, \ldots, M$ sites by defining the multisite occurrence (Wilks, 2009)

$$\chi_t(m) = \begin{cases} 1, & \text{if } w_t(m) \leq \Phi^{-1}\left[p_{ij0}(t)\right]; \\ 0, & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\Phi^{-1}[\cdot]$ is the probit function and $w_t(n) \sim \mathcal{N}[0, 1]$ is Gaussian white noise. To achieve spatially coherent precipitation occurrence, the Markov chain model

in equation (3.7) is forced by a vector of mutually correlated standard Gaussian variates $\vec{w}_t$ characterized by correlation matrix C_R. We populate C_R so that the synthetic correlation matrix C_χ matches its observed counterpart C_x. We achieve this via brute force iteration (Brissette et al., 2007)

$$C_R(i+1) = C_R(i) + \eta(C_x - C_\chi), \tag{3.8}$$

with initial guess $C_R(1) = C_x$, $\eta = 0.1$ and ~ 30 iterations to achieve 10^{-3} precision.

3.5.2 Formulation and Parameter Estimation for Precipitation Amount

In addition, we model precipitation amount using a mixed exponential distribution following Wilks (1999b). The associated probability density function

$$f[r(m)] = \frac{\alpha}{\beta_2(m)} \exp\left[\frac{-r(m)}{\beta_2(m)}\right] + \frac{1-\alpha}{\beta_1(m)} \exp\left[\frac{-r(m)}{\beta_1(m)}\right] \tag{3.9}$$

is the sum of two exponential distributions–the first with larger mean β_2 occurs with probability α, and the second with smaller mean β_1 occurs with probability $1-\alpha$. When precipitation occurs at site m, we choose between β_1 and β_2 according to

$$\beta_t(m) = \begin{cases} \beta_2(m), & \text{if } \frac{\Phi[w_t(m)]}{p_{ij1}(m)} \leq \alpha; \\ \beta_1(m), & \text{otherwise,} \end{cases} \tag{3.10}$$

where $p_{ij1}(m)$ is the appropriate transition probability from (3.6). The formulation in (3.10) captures the tendency for larger precipitation amounts to occur near the interior of wet areas because, for stations and days with small $w_t(m)$ (i.e., first line of (3.10)), other stations around the site are likely to be wet because of the spatial autocorrelation in C_R, and the larger precipitation mean (β_2) is selected (Wilks, 1999a,b). Spatiotemporal variations in these amount parameters for the study region are shown in Fig. 3.6, illustrating that the mixed exponential means (β_1, β_2) tend to be larger in the mountains and outside of summer, and the probability of selecting the larger mean (α) tends to maximize in spring and minimize in summer.

The amount is then recovered from the probability density function via

$$r_t(m) = h - \beta_t(m) \ln[\nu(m)], \tag{3.11}$$

where h is the precipitation occurrence threshold (defined here as 0.254 mm) and $\nu(m) \sim U[0,1]$ is uniformly distributed. The required spatial correlation for $\nu(m)$ is achieved via brute force iteration (Brissette et al., 2007) analogously to the determination of C_R as described prior regarding multisite occurrence. Similar to the probability of precipitation, the means β_1 and β_2 increase markedly at the valley-to-mountain transition near site 15 (Fig. 3.6c-d), and they tend to be smaller in the summer months. The probability of choosing the larger mean (α) is smallest in the summer months for all sites along the transect (Fig. 3.6b).

3.5.3 Formulation and Parameter Estimation for Climate Perturbation

We use the formulation for simulating precipitation occurrence introduced in Smith et al. (2017) where we define perturbed versions of the p_{ijk} values that incorporate trends and sensitivity to oceanic forcing from climate modes such as ENSO and the PDO, extending ideas presented in Woolhiser (2008). In addition, the larger mean in the mixed exponential amount formulation (β_2) is here allowed to have a trend and dependence on oceanic forcing, analogous to the perturbed formulation of p_{ij1}. The perturbed p_{ij1} values are given by

$$p'_{ij1}(t) = p_{ij1}(t) + \gamma_0^{ij1} + \gamma_1^{ij1} t + \gamma_2^{ij1} E(t - \tau_E) + \gamma_3^{ij1} P(t - \tau_P), \tag{3.12}$$

where the $\gamma_{0,1}$ coefficients enable a trend, and the $\gamma_{2,3}$ coefficients provide potentially time-lagged sensitivity to climate variability modes, here chosen to be ENSO (E) and PDO (P) because of their importance to precipitation variability in the example study region (e.g., Wise, 2010). The perturbation of β_2 is formulated analogously. The γ parameters in (3.12) are determined via maximum likelihood in a stepwise fashion, first bringing in the trend, then adding the first oceanic mode, and finally adding the second oceanic mode. At each step, we use the Akaike information criterion to include only parameters that significantly improve the log-likelihood.

As an illustrative example, the perturbed p'_{ij1} values for KSLC over the analysis period (1950-2100) are shown in Figs. 3.7b,c, and an increase in p_{111} is visible over the record. In addition, there is salient periodic variability in all the p_{ij1} time series following the oceanic forcing terms (compare Fig. 3.7b,c to Fig. 3.7a). Determining the perturbation of β_2 via maximum likelihood formulated analogously to (3.12) yielded sensitivity to the oceanic modes with a clear increasing trend over the record (Fig. 3.7d).

We simulated multisite daily precipitation 500 times from 1950 to 2100 to illustrate variability in total precipitation from year to year. Fig. 3.8 shows this variability at KSLC in comparison with the training data. Note the overall increasing trend and low-frequency variability due to ENSO and the PDO. The tendency for correlation between the training data and the ensemble mean arises because the oceanic modes (E and P) driving the precipitation occurrence and amount were diagnosed from the coupled global climate model simulation that produced the training data.

3.6 Discussion and Conclusions

We extended the stochastic temperature simulation framework introduced by Smith et al. (2017) by generalizing the single-temperature, single-site formulation to encompass maximum and minimum temperatures correlated between multiple sites. In addition, we presented a compatible multisite daily precipitation simulation framework based on Markov chain ideas introduced by Woolhiser (2008) and (Wilks, 1998, 2009). The precipitation framework can capture lagged dependence on climate modes such as ENSO and PDO, and was generalized here to capture trends associated with climate change. A transect through complex terrain in northern Utah was used to illustrate spatiotemporal variations in the model parameters, including their trends.

One key difference between the mathematical formulation of single-site SHArP introduced in Smith et al. (2017) is the change from temporally-varying noise coefficient vectors (one for dry days and one for wet days) to noise coefficient matrices that depend on the multisite spatial pattern of precipitation and simulate

observed intersite correlations in temperature. M sites yields an unmanageably large 2^M possible precipitation spatial patterns, so we employ empirical orthogonal function analysis to reduce the number of possible precipitation occurrence patterns to some number much smaller than 2^M. Here, we used the leading two empirical orthogonal functions for illustration, with the first capturing the contrast between all sites wet versus all sites dry, and the second capturing mountain wet / valley dry versus valley dry / mountain wet. The number of EOFs used might be increased depending on the patterns of variability in a particular study region, but needs to be balanced against the accompanying decrease in sample size for estimation of the covariance matrices. After the residual error for each day is assigned to one of four noise coefficient ($\mathbf{C}_k$) matrices, the entries in the matrices are temporally smoothed via Fourier analysis, and those curves are used in the generation process. Naturally, this method removes some of the details, but we found that two harmonics are sufficient for capturing the statistical properties of the input data, and use of more harmonics changed the results minimally.

Another key change associated with the multisite generalization is related to the $\mathbf{A}$ matrix in (3.1), which replaces the scalar a in the single-site, single-temperature case. $\mathbf{A}$ contributes to the autocorrelation of maximum temperature (T_x) and minimum temperature (T_n) at each site, and is block diagonal so it provides no intersite effects. $\mathbf{A}$ more specifically has four entries for each site: the dependence of T_x on the previous day's T_x and T_n, and the dependence of T_n on the previous day's T_n and T_x. The dependence of T_x on the previous day's T_n is arguably the least physical of the four relationships, and could be omitted if weak. The block diagonal assumption on $\mathbf{A}$ means that all intersite correlation is handled by the noise $\mathbf{C}_k$ matrices described above, and is consistent with block diagonal assumptions made for the matrices that control autocorrelation of temperature noise in versions of the Richardson model presented in previous studies (e.g., Richardson, 1981; Wilks, 1999b, 2009).

Testing the encoding of SHArP, we verified that the model accurately estimates the parameters of a broad range of synthetic multisite input data we generated using (3.1) (i.e., the estimation procedures recover $\mathbf{A}_k$, $\mathbf{B}_k$, and $\mathbf{C}_k$). Tested examples

include the full model (3.1) and also simplifications such as vector autoregressive process ($\mathbf{B}_k = 0$) with temporally invariant $\mathbf{C}_k$, or $\mathbf{C}_k$ matrices populated with smooth harmonic time series. Diagnostics such as intersite correlation matrices are also skillfully recovered by SHArP when the training data are generated consistent with its model formulation. Similar to other weather generators, though, the performance diagnostics of SHArP can of course degrade for processes with prominent components (e.g., nonlinearities) not captured by its basic formulation. As an example, SHArP currently assumes $\mathbf{A}_k$ does not depend on time, so generating synthetic training data using (1) with marked temporal fluctuations in $\mathbf{A}_k$ produces positive or negative discrepancies in intersite correlation at different times during the annual cycle. The actual training data used here (statistically downscaled temperatures) resemble observations by design, and certainly contain variability associated with processes not in the SHArP formulations. Nonetheless, fitting SHArP to these data yields intersite squared correlations that match the corresponding training data diagnostics to within 0.05, suggesting that the formulation in (3.1) is sufficiently complex, while still parsimonious.

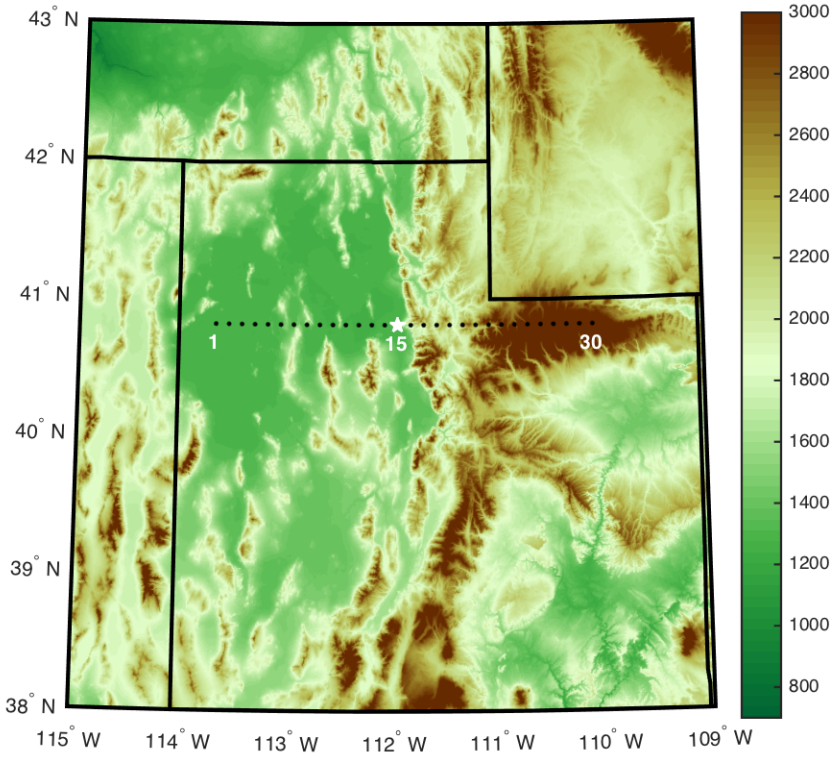

Figure 3.1. Domain map showing northern and central Utah and the transect of 30 sites from the west desert of Utah to the Uinta Mountains. Half of the sites are in the "valley" region (sites 1-15), and half of the sites are located in the mountainous region (Wasatch and Uinta Mountains; sites 16-30). The transect crosses the point nearest the Salt Lake International Airport (KSLC; site 15 indicated by the white star). Color shading indicates elevation in meters above sea level.

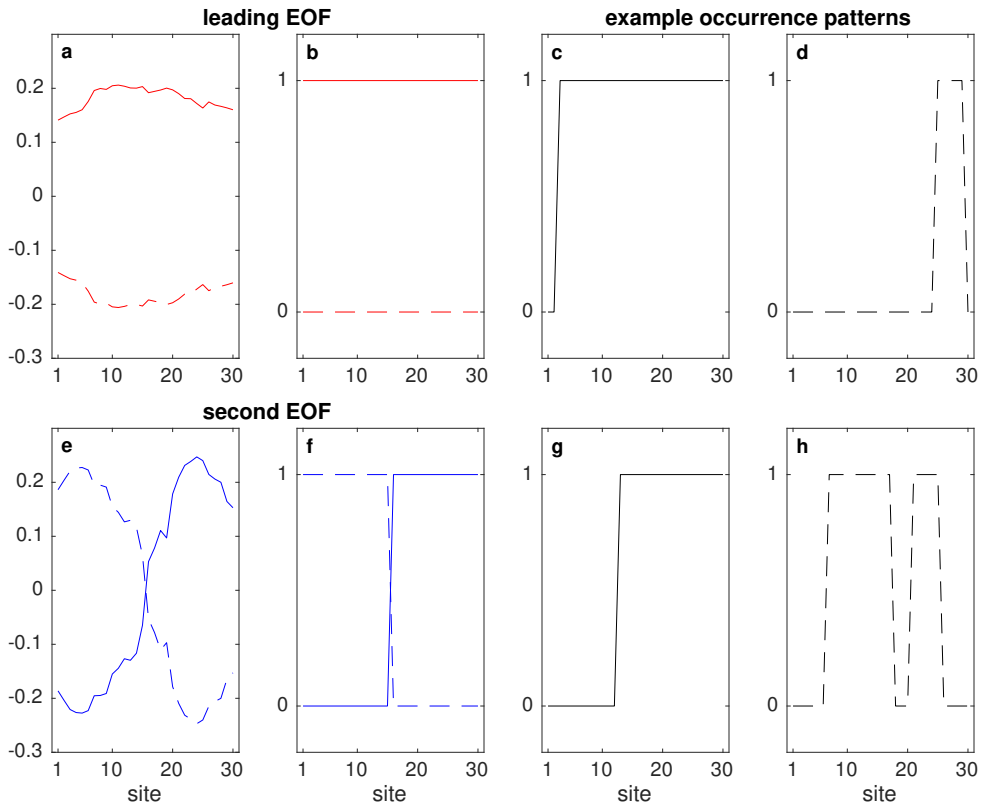

Figure 3.2. Empirical orthogonal functions (EOFs) of precipitation occurrence along the transect. (a) The leading EOF (all sites are wet or all sites are dry), (b) the quantized version of the leading EOF, and example days categorized as the (c) positive polarity and (d) negative polarity of the leading EOF. (e-h) Same as (a-d), but for the second EOF, which captures mountain sites wet / valley sites dry in its positive polarity.

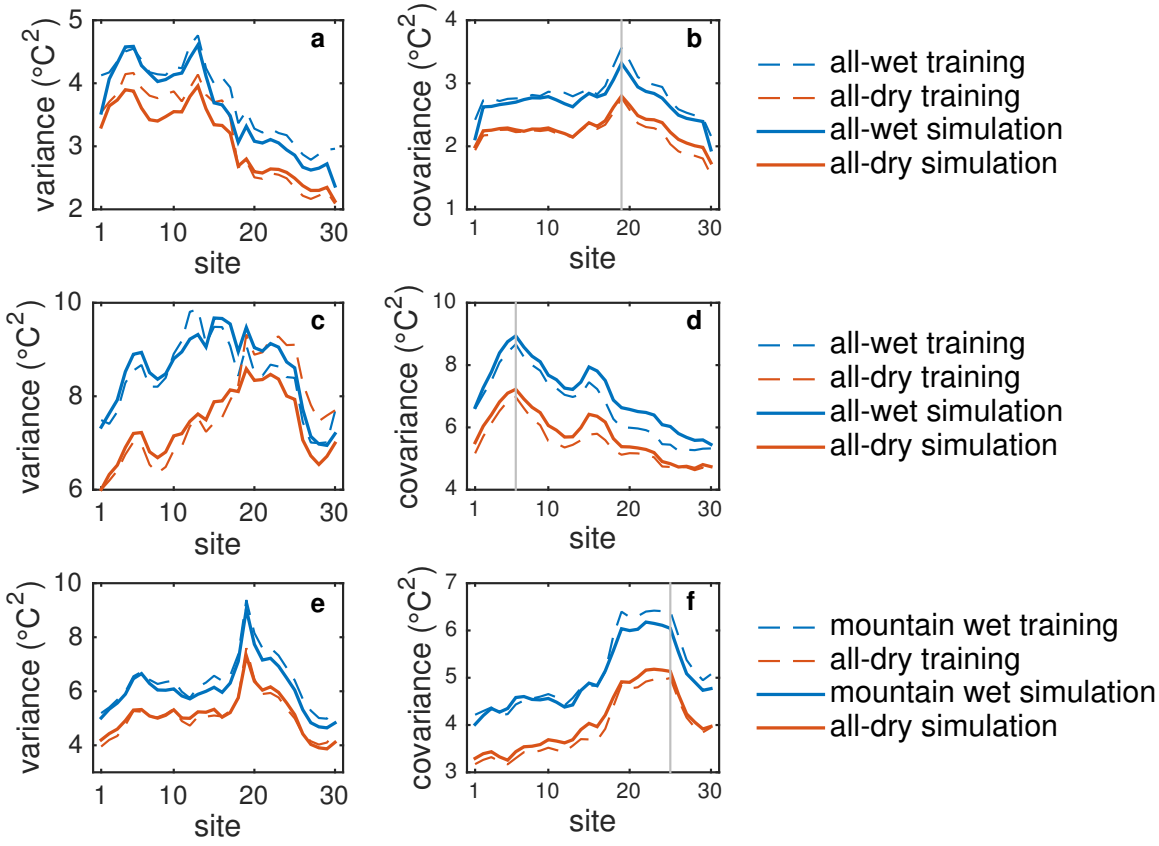

Figure 3.3. (a) Composite variance of the maximum temperature stochastic residuals ($\mathbf{T}_{k+1} - \mathbf{A}\mathbf{T}_k - \mathbf{B}_k$) for days in June 1950-2100 that were all-wet (i.e., positive polarity of EOF 1) or all-dry (i.e., negative polarity of EOF 1). (b) Same as (a), but composite covariance between each site and the site indicated by vertical gray line. (c,d) Same as (a,b) but for December. (e,f) Same as (a,b) but for minimum temperature stochastic residuals for days in October that were mountain wet / valley dry (i.e., positive polarity of EOF 2) or all-dry. In all panels, results from training data are dashed and results from the SHArP simulation are solid.

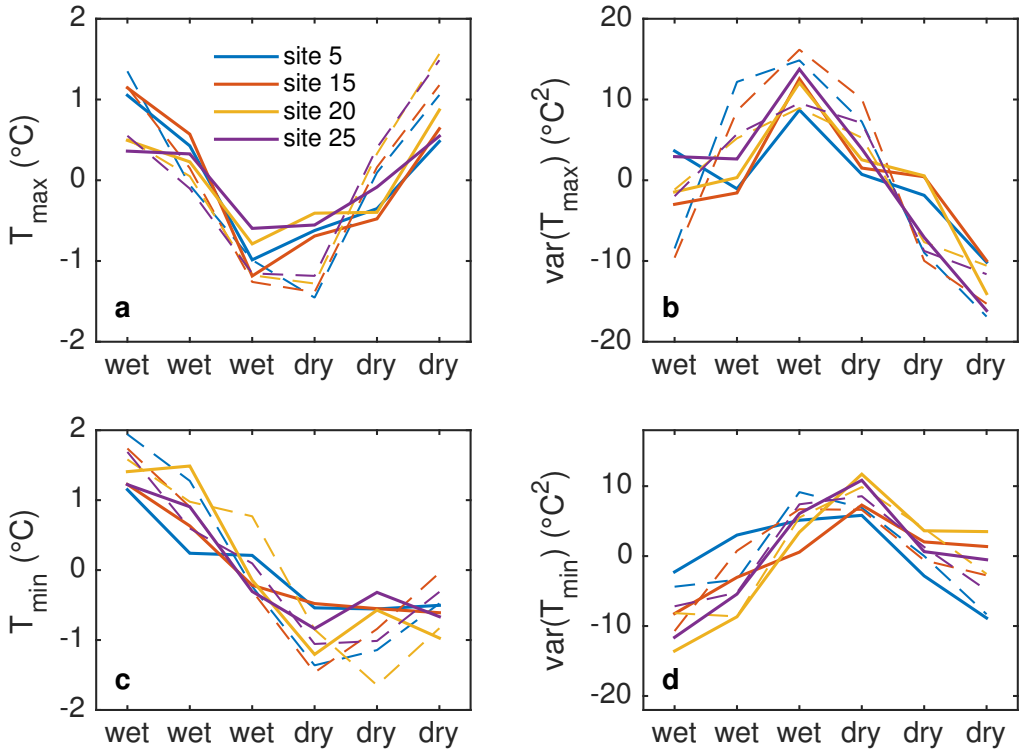

Figure 3.4. Composite temperature evolution for sequences of three all-wet days (i.e., positive polarity of EOF 1) followed by three dry days (i.e., negative polarity of EOF 1) during July-September 1950-2100 at four sites. Plotted values are (a) maximum air temperature, (b) variance of maximum air temperature, (c) minimum air temperature, and (d) variance of minimum air temperature. All composite time series were centered to facilitate comparison of amplitudes, training data are dashed, and the SHArP simulation data are solid.

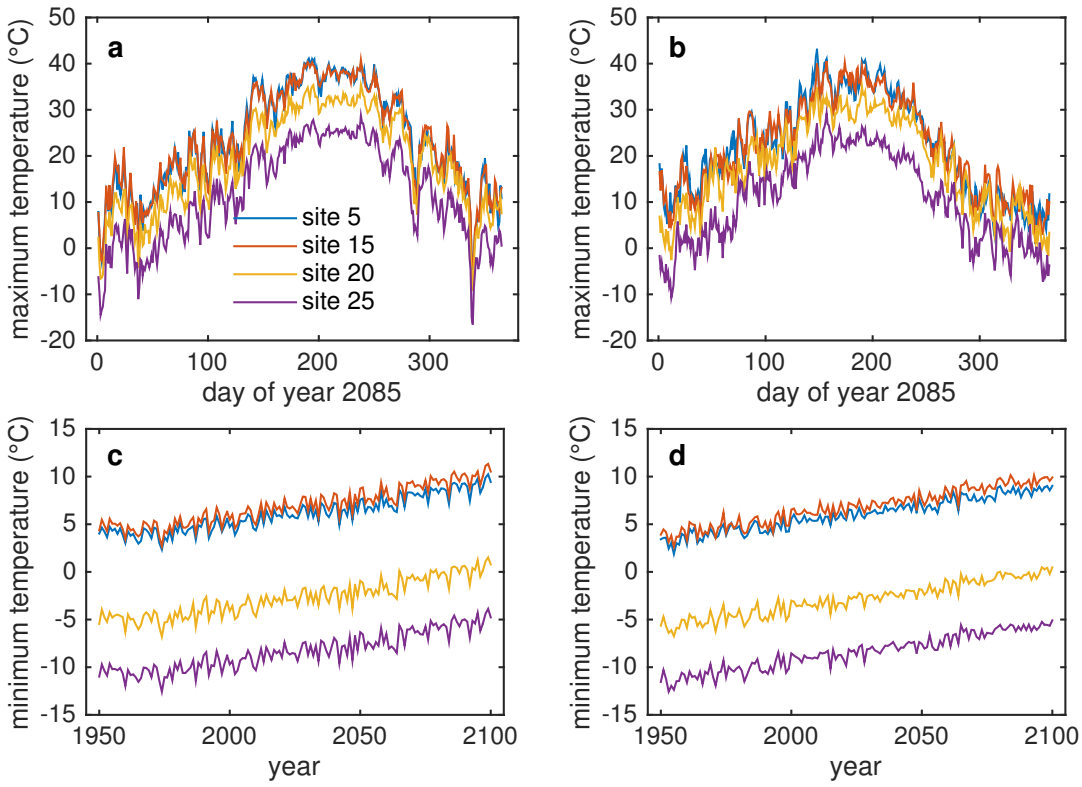

Figure 3.5. Daily maximum temperature in 2085 at four sites for (a) the training data and (b) a sample realization from SHArP. Annual mean minimum temperature from (c) training data and (d) SHArP at the same four sites.

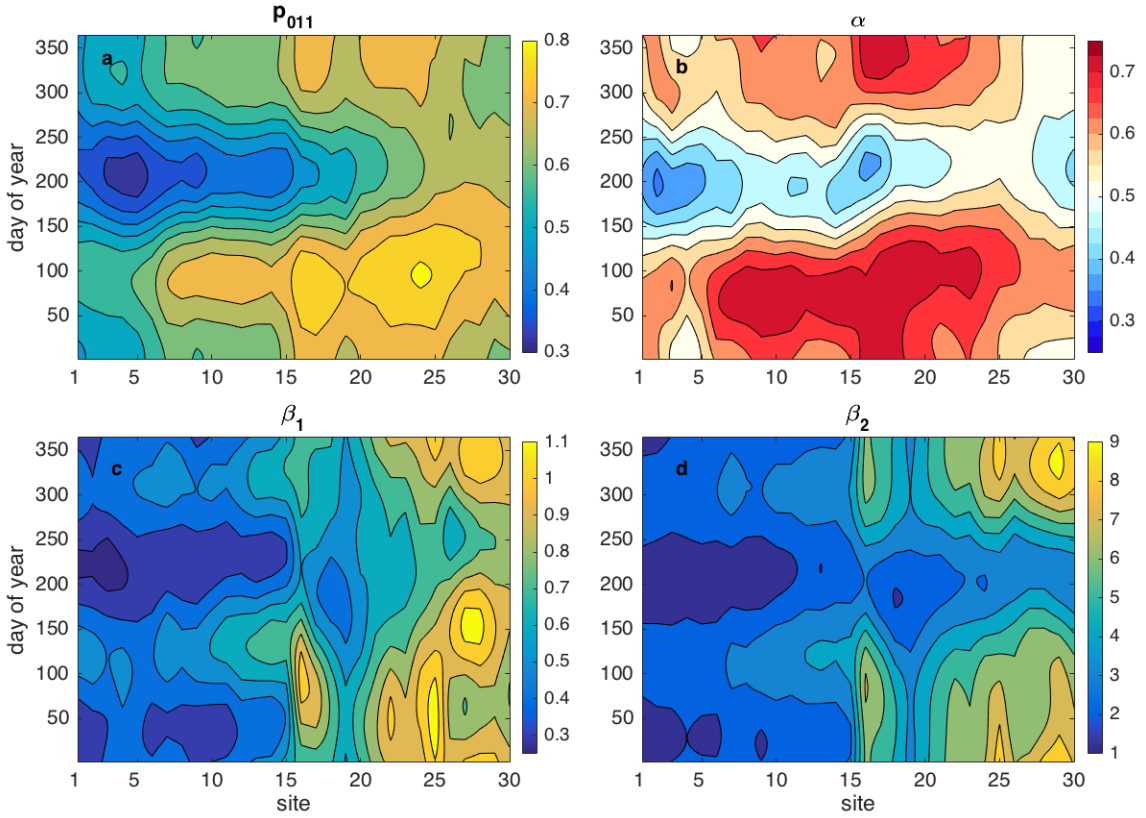

Figure 3.6. For sites 1-30 along the study transect: (a) raw (nonperturbed) probability of precipitation given that the preceding two days were dry and wet, respectively, (b) the probability of selecting the higher precipitation mean from the mixed exponential precipitation distribution (α), (c) the lower mean from the mixed exponential precipitation amount distribution (β_1; units are mm), and (d) the raw (nonperturbed) higher mean from the mixed exponential precipitation amount (β_2; units are mm); Note the different scales for β_1 and β_2.

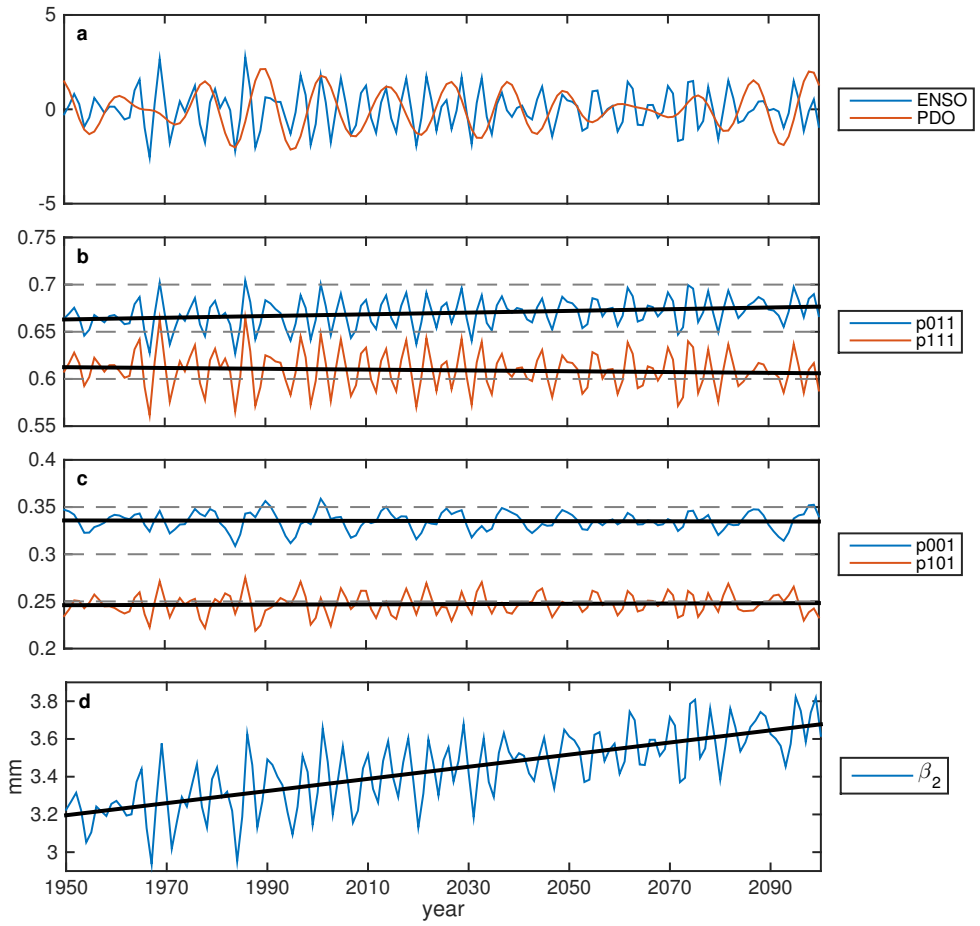

Figure 3.7. (a) Standardized indices of the oceanic modes of variability (ENSO and PDO). (b,c) Annual mean perturbed p_{ij1} values for KSLC with trend lines indicated in black.(d) Annual mean perturbed β_2 values per year for KSLC.

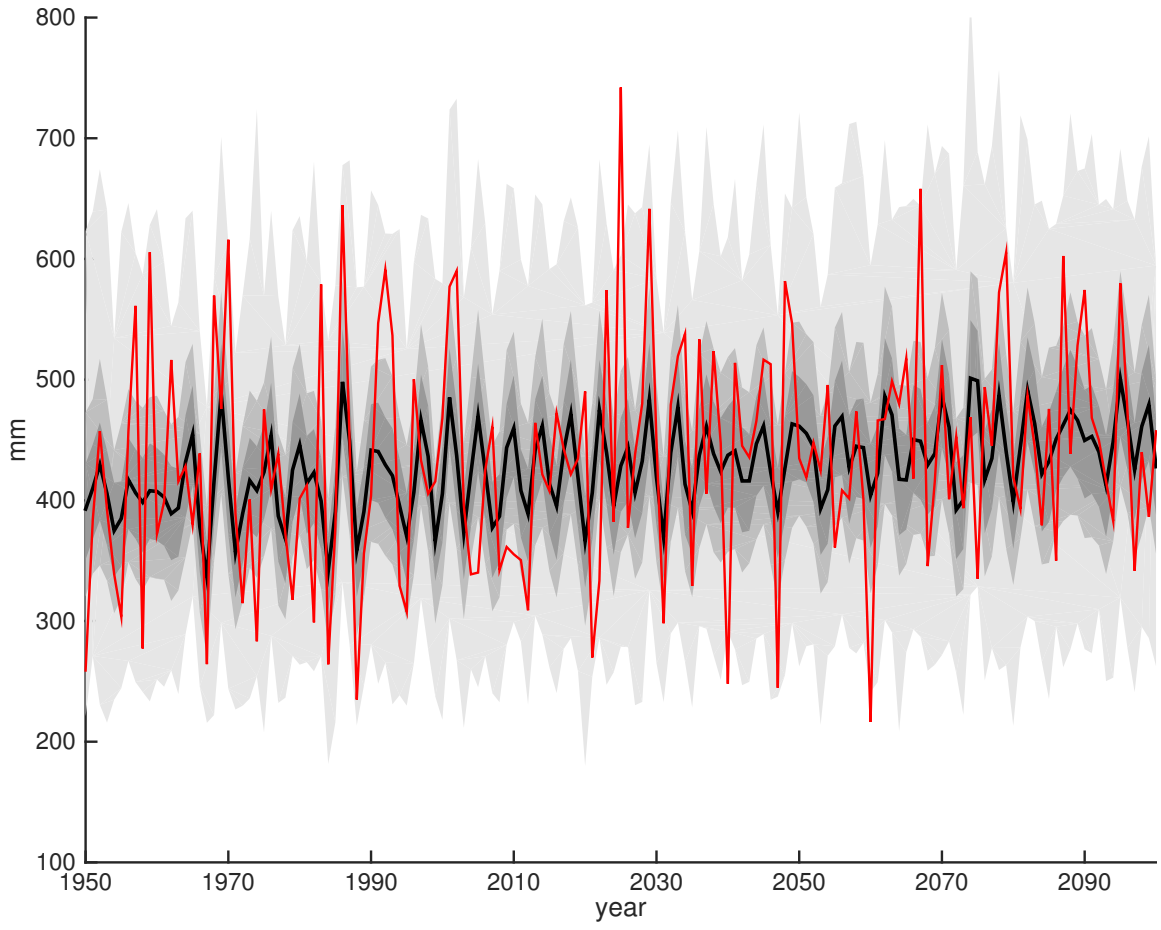

Figure 3.8. Annual total precipitation for KSLC over the period 1950-2100. The mean of the data is shown by the solid black line; the 25th and 75th percentiles, the 10th and 90th percentiles, and the max and min are shaded gray. The total number of simulations is 500. The training data from BCCA CCSM4 are shown in red.

3.7 References

Brissette, F., M. Khalili, and R. Leconte, 2007: Efficient stochastic generation of multi-site synthetic precipitation data. *Journal of Hydrology*, **345 (3–4)**, 121 – 133, doi:10.1016/j.jhydrol.2007.06.035, URL http://www.sciencedirect.com/science/article/pii/S002216940700385X.

Caraway, N. M., J. L. McCreight, and B. Rajagopalan, 2014: Multisite stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *Journal of Hydrology*, **508**, 197 – 213, doi:http://dx.doi.org/10.1016/j.jhydrol.2013.10.054, URL http://www.sciencedirect.com/science/article/pii/S0022169413007981.

Chen, J., F. Brissette, and X. Zhang, 2014: A multi-site stochastic weather generator for daily precipitation and temperature. *Transactions of the ASABE*, **57 (5)**, 1375–1391.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, **27 (9)**, 1119–1152, doi:10.1002/joc.1499, URL http://dx.doi.org/10.1002/joc.1499.

Kleiber, W., R. W. Katz, and B. Rajagopalan, 2012: Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. *Water Resources Research*, **48 (1)**, n/a–n/a, doi:10.1029/2011WR011105, URL http://dx.doi.org/10.1029/2011WR011105, w01523.

Kleiber, W., R. W. Katz, and B. Rajagopalan, 2013: Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, **7 (1)**, 588–612, doi:10.1214/12-AOAS602, URL http://dx.doi.org/10.1214/12-AOAS602.

Maurer, E. P., L. Brekke, T. Pruitt, and P. B. Duffy, 2007: Fine-resolution climate projections enhance regional climate change impact studies. *Eos, Transactions American Geophysical Union*, **88 (47)**, 504–504, doi:10.1029/2007EO470006, URL http://dx.doi.org/10.1029/2007EO470006.

Reclamation, 2013: Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of downscaled CMIP5 climate projections, comparion with preceding information, and summary of user needs. Tech. rep., U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, Colorado. 47pp.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17 (1)**, 182–190, doi:10.1029/WR017i001p00182, URL http://dx.doi.org/10.1029/WR017i001p00182.

Smith, K., C. Strong, and F. Rassoul-Agha, 2017: A new method for generating stochastic simulations of daily air temperature for use in weather generators. *Journal of Applied Meteorology and Climatology*, **56 (4)**, 953–963, doi:10.1175/JAMC-

D-16-0122.1, URL http://dx.doi.org/10.1175/JAMC-D-16-0122.1, http://dx.doi.org/10.1175/JAMC-D-16-0122.1.

Smith, K., C. Strong, and S.-Y. Wang, 2015: Connectivity between historical Great Basin precipitation and Pacific Ocean variability: A CMIP5 model evaluation. *Journal of Climate*, **28 (15)**, 6096–6112, doi:10.1175/JCLI-D-14-00488.1, URL http://dx.doi.org/10.1175/JCLI-D-14-00488.1.

Stern, R. D. and R. Coe, 1984: A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, **147 (1)**, 1–34, URL http://www.jstor.org/stable/2981736.

Thompson, G. A. and D. B. Burke, 1974: Regional geophysics of the basin and range province. *Annual Review of Earth and Planetary Sciences*, **2**, 213–238.

Todorovic, P. and D. A. Woolhiser, 1975: A stochastic model of ω-day precipitation. *Journal of Applied Meteorology*, **14**, 17–24, doi:10.1175/1520-0450(1975)014⟨0017:ASMODP⟩2.0.CO;2, URL http://dx.doi.org/10.1175/1520-0450(1975)014<0017:ASMODP>2.0.CO;2.

Wilks, D., 1998: Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, **210 (1–4)**, 178 – 191, doi:10.1016/S0022-1694(98)00186-3, URL http://www.sciencedirect.com/science/article/pii/S0022169498001863.

Wilks, D., 1999a: Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, **93 (3)**, 153–169, URL http://www.sciencedirect.com/science/article/pii/S0168192398001257.

Wilks, D., 1999b: Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, **96 (1–3)**, 85 – 101, doi:10.1016/S0168-1923(99)00037-4, URL http://www.sciencedirect.com/science/article/pii/S0168192399000374.

Wilks, D. S., 2008: High-resolution spatial interpolation of weather generator parameters using local weighted regressions. *Agricultural and Forest Meteorology*, **148 (1)**, 111 – 120, doi:10.1016/j.agrformet.2007.09.005, URL http://www.sciencedirect.com/science/article/pii/S0168192307002511.

Wilks, D. S., 2009: A gridded multisite weather generator and synchronization to observed weather data. *Water Resources Research*, **45 (10)**, n/a–n/a, doi:10.1029/2009WR007902, URL http://dx.doi.org/10.1029/2009WR007902.

Wise, E. K., 2010: Spatiotemporal variability of the precipitation dipole transition zone in the western United States. *Geophysical Research Letters*, **37 (7)**, doi:10.1029/2009GL042193, URL http://dx.doi.org/10.1029/2009GL042193.

Woolhiser, D. A., 2008: Combined effects of the Southern Oscillation Index and the Pacific Decadal Oscillation on a stochastic daily precipitation model. *Journal of Climate*, **21 (5)**, 1139–1152, doi:10.1175/2007JCLI1862.1, URL http://dx.doi.org/10.1175/2007JCLI1862.1.

CHAPTER 4

USING THE METHOD OF LARGE DEVIATIONS TO SIMULATE EXTREME PRECIPITATION SEQUENCES

4.1 Abstract

A limitation of traditional stochastic weather generators is their ability to capture meteorological extremes, including the occurrence of dry or wet spells. While second-order Markov chains have been found to better produce the dry spells that are common in the western U.S., one would need many thousands of years of generated data in order to find a few years that are considered "extreme" yet still statistically consistent with the training data. In order to avoid this so-called "wait to get lucky" method, the probabilities of precipitation are modified using the method of large deviations. This mathematically-based method is shown to accurately modify the probabilities of precipitation so as to produce a variety of specified precipitation occurrence sequences that are extreme yet consistent with the statistical properties of the underlying training data. The method is illustrated for the Salt Lake International Airport (KSLC), a site within the Great Basin in the western U.S.

4.2 Introduction

Stochastic weather generators (SWGs) are primarily used as an alternative to low-resolution global climate models (GCMs); they are able to statistically match the input data (commonly either observations or GCM output) and produce high-resolution output on a point scale. They are especially useful tools in areas that have missing or a lack of meteorological data. An important application of SWGs

is their ability to capture meteorological extremes that are statistically consistent with, but may not have occurred in, the input data (e.g., long-term droughts). This application will become increasingly important in the future as the climate continues to change due to anthropogenic forcing. Hydrologists and water managers may need rapidly developed stochastic ensembles to capture uncertainty associated with, for example, changes in snowpack due to drought and more winter precipitation falling as rain rather than snow in the higher elevations due to warming temperatures.

Most existing studies that evaluate SWGs on their ability to capture and generate extremes pertain to high-intensity precipitation events that lead to flooding or hot/cold temperature spells. Vrac and Naveau (2007) used extreme value theory to determine that a combination of gamma and generalized Pareto distributions best capture all precipitation events (rather than just a gamma distribution, which does best at capturing low to moderate intensities). Furrer and Katz (2008) discuss how existing parametric SWGs, including the generalized linear models (GLMs), are unable to capture the very high intensity precipitation amounts with their amount distributions. GLMs are a type of SWG that can more easily model discrete variables and variables with non-normal distributions (McCullagh and Nelder, 1989). Their study also found that a combination of a gamma distribution for low-to-moderate precipitation intensities and a generalized Pareto distribution for much higher intensities considerably improves the ability of the SWG to capture the very high intensity events that are uncommon but still present in nature.

Wilks (1999) studied the effects of different Markov chain models and different precipitation amount distributions on reproducing dry/wet spells and extreme precipitation intensities. The hybrid-order Markov chains were best able to capture extended droughts, especially in the western U.S. The Gamma distribution for precipitation amounts was unable to capture extreme intensities while the mixed exponential distribution better captured extremes but for areas where the daily amounts are lower (50-100 mm). Other studies have analyzed the ability of SWGs to capture hot and cold temperature spells. Kyselý and Dubrovský (2005) evaluated a SWG on its ability to capture extreme temperature events in Europe. The au-

thors found that the improvements made to the SWG to capture extremes did not do well at capturing daily extreme maximum or minimum temperatures, but the addition of a monthly weather generator with intermonthly variability improved the ability of the SWG to simulate heat waves.

There is a lack of existing literature discussing the ability of SWGs to capture extremes in a Markov chain-based precipitation occurrence process (i.e., dry or wet spells, not necessarily high-intensity rainfall events). It is important for SWGs to be able to generate extremes, especially in semi-arid regions that rely on snowpack-dominated watersheds (i.e., the Great Basin in the western U.S.) as the climate continues to change. In this study, we use the mathematical method of large deviations to modify the probabilities of precipitation that allow for various dry (or wet) spells. This method is ideal because we do not have to "wait to get lucky" by simulating thousands of years of precipitation occurrences and only pulling out the "extreme" years. Here, we present the mathematics behind the method and illustrate the changes in precipitation occurrence with the new probabilities of precipitation that yield "extreme" events.

4.3 Data and Study Area

The SWG input data used in this study is 0.125° bias corrected constructed analogs (BCCA) of daily CCSM4 output from the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Maurer et al., 2007; Reclamation, 2013). For this analysis, we used the historical output, which spans from 1 January 1950 to 31 December 2005, in addition to the future RCP 8.5 (high emissions scenario) output, which spans from 1 January 2006 to 31 December 2100. For this study, we specifically focused on the Great Basin of the western U.S., known for its basin-and-range topography (Thompson and Burke, 1974), because it deals with recurring drought and is susceptible to more instances of drought as the climate warms. In addition, the previous stochastic harmonic autoregressive parametric (SHArP) weather generator studies used the region as a study area (Smith et al., 2017). We selected the site closest to the Salt Lake International Airport (KSLC; 40.8125°N, 111.9375°W) for illustration.

4.4 Method of Large Deviations

We start with the raw, nonperturbed p_{ijk} values for a single day of year at a single site (in this case, KSLC) determined from a two-state (wet or dry), second-order Markov chain given by

$$p_{ijk} = P\left\{X_t = k | X_{t-1} = j,\ X_{t-2} = i\right\}.$$

In the equation above, i, j, and k can take value 1 to indicate a wet day [at least 0.254 mm (0.01 inches) of precipitation] or value 0 to indicate a dry day. Previous studies have shown that second-order Markov chains are better than first-order Markov chains at capturing the dry spells that are common in the western U.S. (Stern and Coe, 1984; Wilks, 1999).

The transition matrix $\mathbf{p}$ containing the raw p_{ijk} values is

$$\mathbf{P} = \begin{bmatrix} p & 1-p & 0 & 0 \\ 0 & 0 & q & 1-q \\ r & 1-r & 0 & 0 \\ 0 & 0 & s & 1-s \end{bmatrix},$$

where $p = p_{000}$, $1-p = p_{001}$, $q = p_{010}$, $1-q = p_{011}$, $r = p_{100}$, $1-r = p_{101}$, $s = p_{110}$, and $1-s = p_{111}$. To apply the method of large deviations, we turn the second-order Markov chain into a first-order Markov chain by increasing the state space from $\{0,1\}$ to $\{00,01,10,11\}$. The transition matrix $\mathbf{Q}$ of the extreme process is denoted by

$$\mathbf{Q} = \begin{bmatrix} x & 1-x & 0 & 0 \\ 0 & 0 & y & 1-y \\ z & 1-z & 0 & 0 \\ 0 & 0 & t & 1-t \end{bmatrix}.$$

The invariant measure (or steady state) for this Markov chain is obtained by solving

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \mathbf{Q} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \quad \text{and}$$

$$a + b + c + d = 1\,, \tag{4.1}$$

where

$$\begin{aligned} a &= \frac{1}{1 + (2(1-x)/z) + ((1-x)(1-y)/zt)}\,, \\ b &= \frac{1}{2 + (z/(1-x)) + ((1-y)/t)}\,, \\ c &= \frac{1}{2 + (z/(1-x)) + ((1-y)/t)}\,, \\ d &= \frac{1}{(zt/[(1-x)(1-y)]) + [2t/(1-y)] + 1}\,. \end{aligned} \tag{4.2}$$

The expected number of 0s in n days is given by $(a+b)n$ and the expected number of 1s is given by $[1-(a+b)]n$ as n goes to infinity. Note that b and c are equal. The entries in $\mathbf{Q}$ give us the new probabilities of precipitation; x gives the new value for p_{000}, y gives the new value for p_{010}, and so on.

By the Gibbs conditioning principle in large deviation theory (Csiszár et al., 1987; Rassoul-Agha and Seppäläinen, 2015, Thm. 4, Thm. 13.5, respectively), the entries of $\mathbf{Q}$ are obtained by minimizing the entropy H given by

$$\begin{aligned} H(x,y,z,t) = {} & ax\log(x/p) + a(1-x)\log[(1-x)/(1-p)] \\ & + by\log(y/q) + b(1-y)\log[(1-y)/(1-q)] \\ & + cz\log(z/r) + c(1-z)\log[(1-z)/(1-r)] \\ & + dt\log(t/s) + d(1-t)\log[(1-t)/(1-s)]. \end{aligned} \tag{4.3}$$

Recall that a, b, c, d are functions of x, y, z, t as in 4.2.

This minimization is done subject to a given constraint that describes the extreme event we are aiming to generate. We do this by applying the *fmincon.m* Matlab function and setting a nonlinear constraint defining the minimum or maximum percentage of dry days such as

$$a + b \geq 0.9, \tag{4.4}$$

where $a+b$ indicates the proportion of dry days and 0.9 indicates that at least 90% of the total number of days in the simulation are dry. Note that even though 4.4 is

linear, a and b are nonlinear functions of x, y, z, t. We can also define the constraint as the percentage of at least a given number of consecutive dry days indicating dry spells such as

$$ax^{k-2} \geq 0.9, \tag{4.5}$$

where a is the probability of the first two days being dry, x is the probability of the next day being dry given that the previous two days were dry, k indicates the number of consecutive dry days, and 0.9 indicates that at least the given number of consecutive dry days occurs at least 90% of the time. The constraint in 4.5 will include, and likely prefer, precipitation sequences that have more than the specified number of dry days in the constraint. In order to obtain dry spells of exactly k days (where the k number of dry days are bounded by wet days), the constraint is given by

$$czx^{k-2}(1-x) \geq 0.9, \tag{4.6}$$

where c is the probability of the first day being wet and the second day being dry (sequence 10), z is the probability of the sequence 00 given that the previous sequence was 10, x is the probability of the sequence 00 given that the previous sequence was 00, k indicates the number of consecutive dry days, and $1-x$ is the probability of the sequence 01 given that the previous sequence was 00. These constraints could also be made to specify wet days or wet spells. Once minimization is achieved, we take the output as new, modified p_{ijk} values and use the new values to simulate the "extreme" precipitation occurrence.

4.5 Illustrative Simulations

In this study, we focus on dry spells specifically. We simulate many years of data specific to a chosen month, and use the p_{ij0} values near the middle of the month to be representative. Generalization to the more elaborate case with temporally varying p_{ij0} functions is discussed in Section 4.6.

Table 4.1 compares the p_{ij0} values between the training data (on day of year 75) and the "extreme" simulation data with varying constraints. Fig. 4.1 shows a sample of 1000 days in the training data (simulated using only the p_{ij0} values from March 16th) and the same 1000 days in a simulation where the number of dry days

in the simulation was set to at least 90%. By visual inspection, the simulation has fewer wet days; in those 1000 days, 583 are wet in the training data and only 170 are wet in the simulation. Figs. 4.2 and 4.3 show the difference between training data and simulated data but for at least two consecutive dry days and at least five consecutive dry days, respectively; Figs. 4.4 and 4.5 also show the difference but for exactly five consecutive dry days and exactly ten consecutive dry days, respectively. Note the much larger number of total dry days as the number of consecutive dry days increases (and the larger number of total dry days for the "at least" constraint versus the "exact" constraint).

It is important to note that the extreme process is still a stochastic process. It has the same statistics as the original process but is conditioned on observing the extreme sequence. As such, it may happen that the extreme process produces its own extreme events such as 50% dry days instead of 90%. Observing the extreme process over a somewhat long period of time will result in 90% dry days. This is similar to how a fair two-sided coin could produce 20% heads but as the number of tosses increases, the proportion approaches 50% heads. Here, the training data and the simulation both span at least 150 years; over the entire simulation, due to the constraint, the total number of dry days should be at least 90%, but this does not necessarily mean that each year or each 1000 days will have at least 90% dry days.

4.6 Discussion and Conclusions

SWGs are useful tools for generating long-term point-scale daily precipitation and air temperature values that statistically match the input data and are presented with the challenge of simulating "extreme" climate scenarios such as droughts, especially in semi-arid regions of complex terrain such as the Great Basin in the western U.S. Instead of "waiting to get lucky" by generating thousands of years of data to get a few instances of "extreme" precipitation occurrence events, such as extended drought, we use the method of large deviations to modify the probabilities of precipitation to give us a specific "extreme" situation. This mathematical method was shown to accurately modify the probabilities of precipitation, result-

ing in binary precipitation occurrence output that matches the "extreme" event of interest.

A limitation of the method of large deviations is the fact that it assumes stationarity because we select the raw, non-perturbed probabilities of precipitation (p_{ijk} values) from one day of year and use only those values to generate the new p_{ijk} values. In reality, the probabilities of precipitation change from day-to-day and even from year-to-year due to oceanic modes of variability such as El Niño–Southern Oscillation or the Pacific Decadal Oscillation. Climate change might also play a role in changing the probabilities of precipitation into the future, particularly if the SWG is forced with future GCM output. Another limitation arises when using the constraint that limits the number of consecutive dry days to a specific number. Because SHArP uses a second-order Markov chain, if the fraction of exactly two consecutive dry days is at least 90%, the method of large deviations limits the resulting precipitation occurrence to the sequence that looks almost exactly like 1001001001 (repeating over the entire simulation) because it is unable to produce three dry days in a row. This is not statistically impossible but it is an unlikely "extreme" event, especially in a semi-arid region such as the Great Basin. Care is needed to make sure the constraint is reasonable for the study area.

In this study, we focused only on modifying the probabilities of precipitation for a single site, and it will be useful to eventually extend this method to account for spatially correlated extremes at multiple sites. One could do this exact process for each site individually, but the resulting spatial correlations may not match the true extreme-event correlation patterns. As it stands, the same basic formula for determining the transition matrix $\mathbf{Q}$ can possibly be used for multiple sites; however, this transition matrix will have a large number of entries due to the addition of multiple sites. The entropy equation H will also depend on these many entries. We are unable to minimize H with the large number of variables in a reasonable amount of time. However, we could impose an additional structure on $\mathbf{Q}$ similar to that of $\mathbf{P}$ for multiple sites to possibly reduce the number of variables and make it reasonable for minimization.

Table 4.1. p_{ij0} values for the training data (on day of year 75) and for the simulated data with various constraints.

	p_{000}	p_{010}	p_{100}	p_{110}
training data	0.6549	0.2780	0.5963	0.2655
fraction of dry days $\geq 90\%$	0.9147	0.4180	0.9170	0.2605
at least 2 cons. dry days $\geq 90\%$	0.9507	0.3164	0.9540	0.2170
at least 5 cons. dry days $\geq 90\%$	0.9812	0.2962	0.8245	0.2461
exactly 5 cons. dry days $\geq 90\%$	0.6535	0.9999	0.9998	0.7780
exactly 10 cons. dry days $\geq 90\%$	0.8216	1.0000	1.0000	0.8215

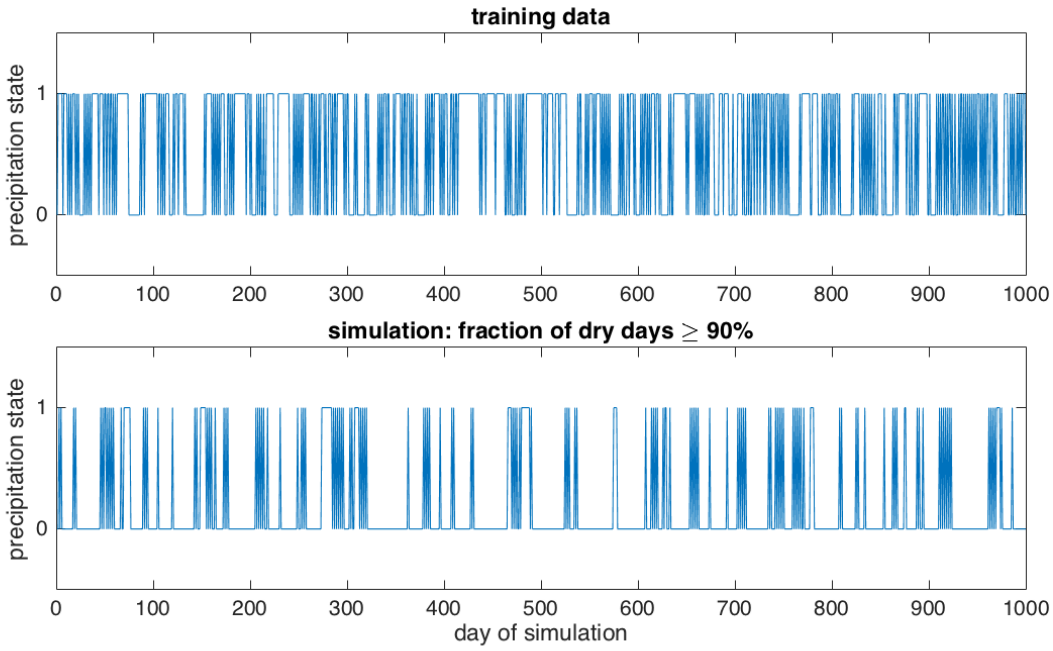

Figure 4.1. Precipitation states (dry = 0, wet = 1) for 1000 days of the training data (top) and a sample simulation (bottom) where the constraint is "fraction of total dry days is at least 90%". In these 1000 days, there are 438 dry days in the training data and 815 dry days in the simulation.

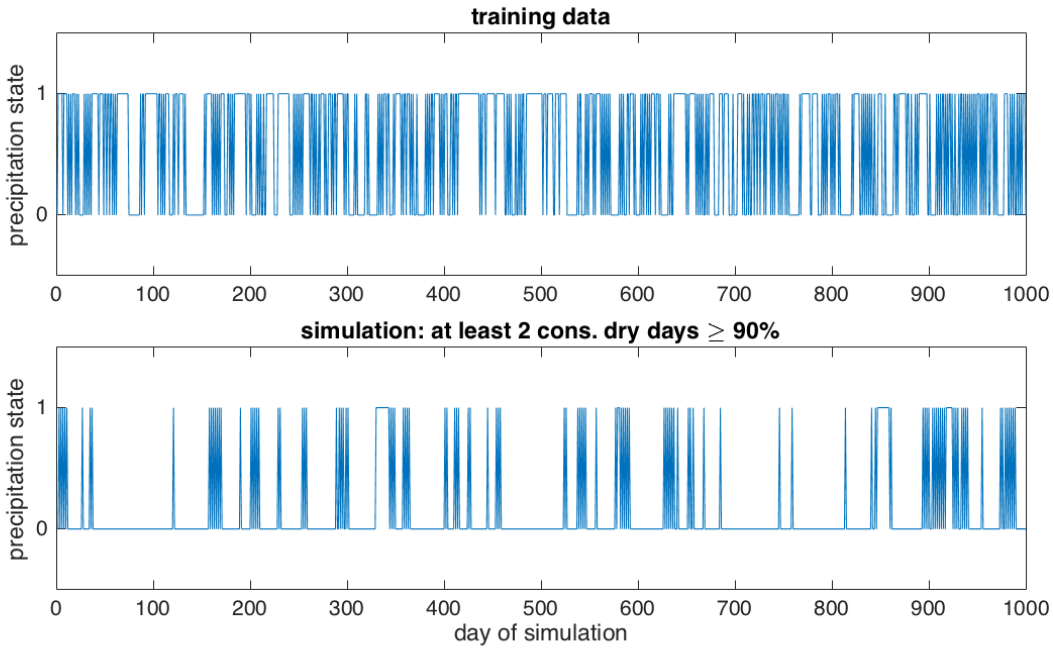

Figure 4.2. Precipitation states (dry = 0, wet = 1) for 1000 days of the training data (top) and a sample simulation (bottom) where the constraint is "at least two consecutive dry days occur at least 90% of the time". In these 1000 days, there are 438 dry days in the training data and 853 dry days in the simulation.

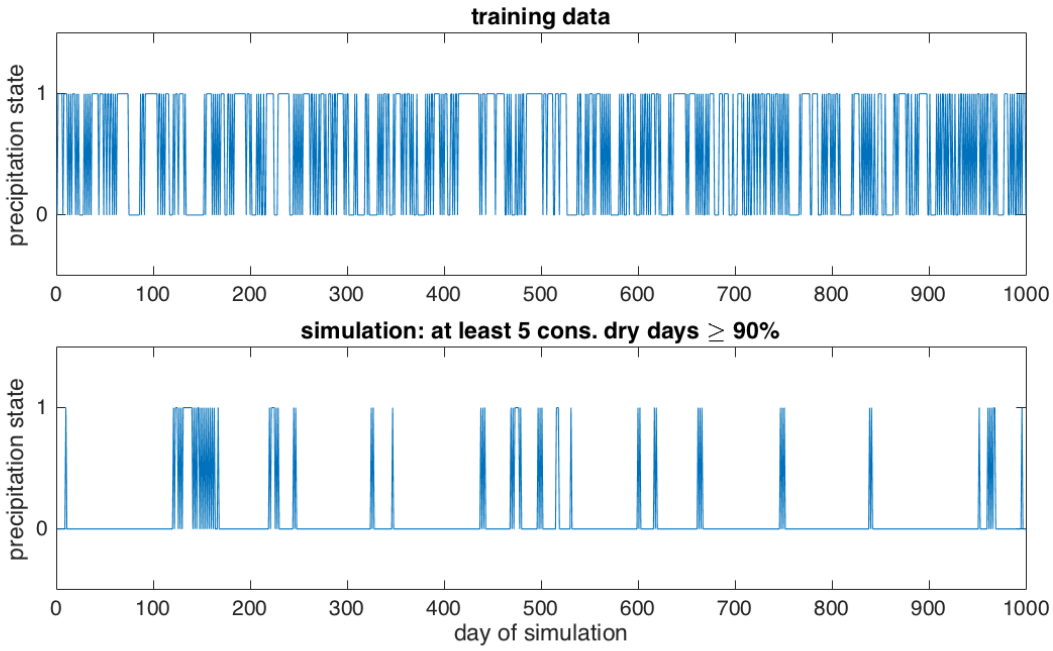

Figure 4.3. Precipitation states (dry = 0, wet = 1) for 1000 days of the training data (top) and a sample simulation (bottom) where the constraint is "at least five consecutive dry days occur at least 90% of the time". In these 1000 days, there are 438 dry days in the training data and 921 dry days in the simulation.

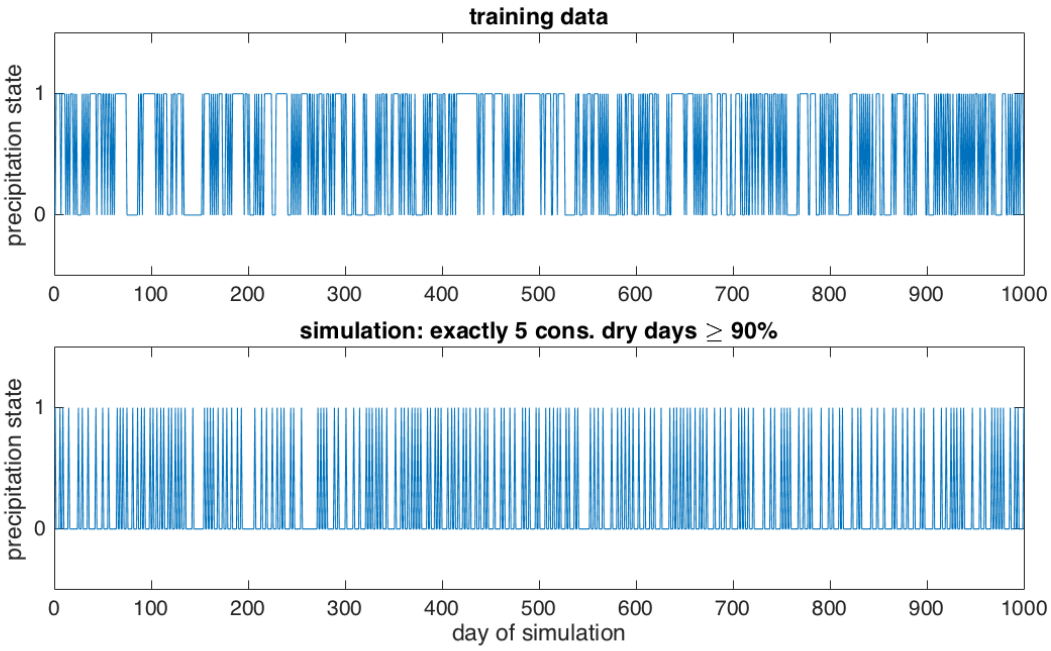

Figure 4.4. Precipitation states (dry = 0, wet = 1) for 1000 days of the training data (top) and a sample simulation (bottom) where the constraint is "exactly five consecutive dry days occur at least 90% of the time". In these 1000 days, there are 438 dry days in the training data and 807 dry days in the simulation.

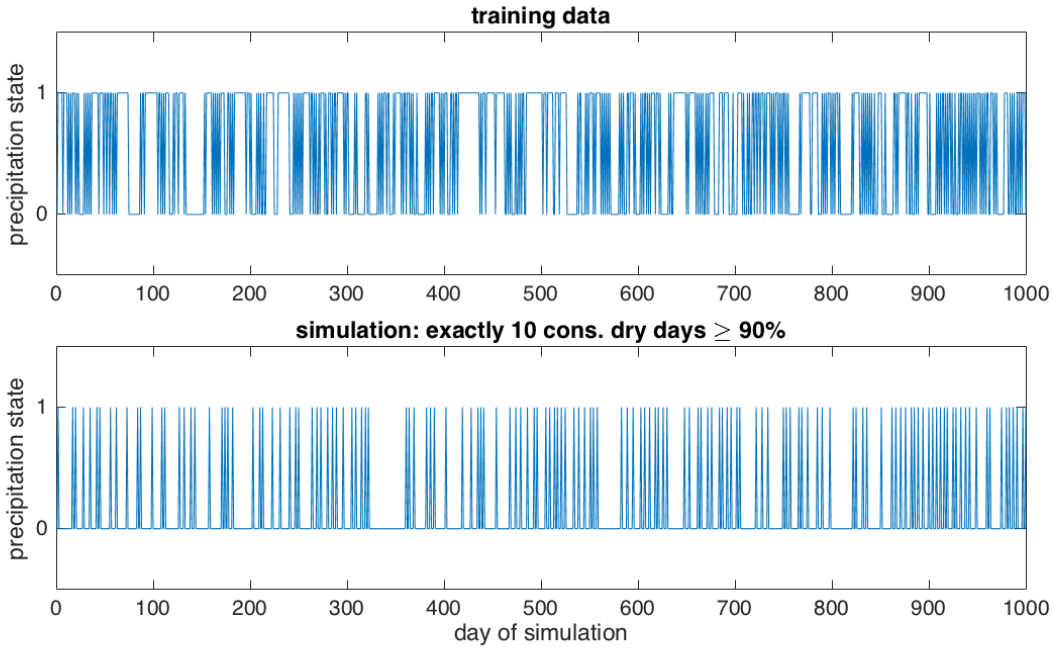

Figure 4.5. Precipitation states (dry = 0, wet = 1) for 1000 days of the training data (top) and a sample simulation (bottom) where the constraint is "exactly ten consecutive dry days occur at least 90% of the time". In these 1000 days, there are 438 dry days in the training data and 857 dry days in the simulation.

4.7 References

Csiszár, I., T. M. Cover, and B. S. Choi, 1987: Conditional limit theorems under Markov conditioning. *IEEE Transactions on Information Theory*, **33 (6)**, 788–801, doi:10.1109/TIT.1987.1057385.

Furrer, E. M. and R. W. Katz, 2008: Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, **44 (12)**, n/a–n/a, doi:10.1029/2008WR007316, URL http://dx.doi.org/10.1029/2008WR007316, w12439.

Kyselý, J. and M. Dubrovský, 2005: Simulation of extreme temperature events by a stochastic weather generator: effects of interdiurnal and interannual variability reproduction. *International Journal of Climatology*, **25 (2)**, 251–269, doi:10.1002/joc.1120, URL http://dx.doi.org/10.1002/joc.1120.

Maurer, E. P., L. Brekke, T. Pruitt, and P. B. Duffy, 2007: Fine-resolution climate projections enhance regional climate change impact studies. *Eos, Transactions American Geophysical Union*, **88 (47)**, 504–504, doi:10.1029/2007EO470006, URL http://dx.doi.org/10.1029/2007EO470006.

McCullagh, P. and J. Nelder, 1989: *Generalized Linear Models*. 2nd ed., Chapman & Hall, London.

Rassoul-Agha, F. and T. Seppäläinen, 2015: *A course on large deviations with an introduction to Gibbs measures*, Graduate Studies in Mathematics, Vol. 162. American Mathematical Society, Providence, RI, xiv+318 pp.

Reclamation, 2013: Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of downscaled CMIP5 climate projections, comparion with preceding information, and summary of user needs. Tech. rep., U.S. Department of the Interior, Bureau of Reclamation, Technical Services Center, Denver, Colorado. 47pp.

Smith, K., C. Strong, and F. Rassoul-Agha, 2017: A new method for generating stochastic simulations of daily air temperature for use in weather generators. *Journal of Applied Meteorology and Climatology*, **56 (4)**, 953–963, doi:10.1175/JAMC-D-16-0122.1, URL http://dx.doi.org/10.1175/JAMC-D-16-0122.1, http://dx.doi.org/10.1175/JAMC-D-16-0122.1.

Stern, R. D. and R. Coe, 1984: A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, **147 (1)**, 1–34, URL http://www.jstor.org/stable/2981736.

Thompson, G. A. and D. B. Burke, 1974: Regional geophysics of the basin and range province. *Annual Review of Earth and Planetary Sciences*, **2**, 213–238.

Vrac, M. and P. Naveau, 2007: Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resources Research*, **43 (7)**, n/a–n/a, doi:10.1029/2006WR005308, URL http://dx.doi.org/10.1029/2006WR005308, w07402.

Wilks, D., 1999: Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, **93 (3)**, 153–169, URL `http://www.sciencedirect.com/science/article/pii/S0168192398001257`.